A.D.E.F
UR4671
Apprentissage | Didactique | Evaluation | Formation

LIS
LABORATOIRE D'INFORMATIQUE & DES SYSTÈMES

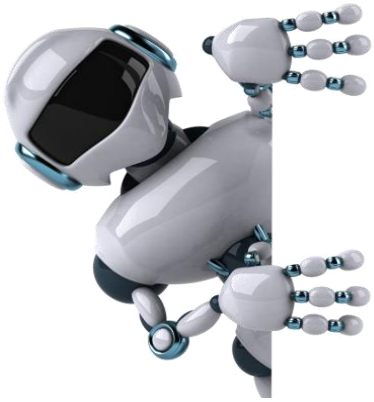
INSPÉ Institut national supérieur du professorat et de l'éducation d'Aix-Marseille

amU Aix Marseille Université

# IA-ÉDU

## Enjeux, Défis et Perspectives

## CORIA-TALN 2025



Actes de l'atelier

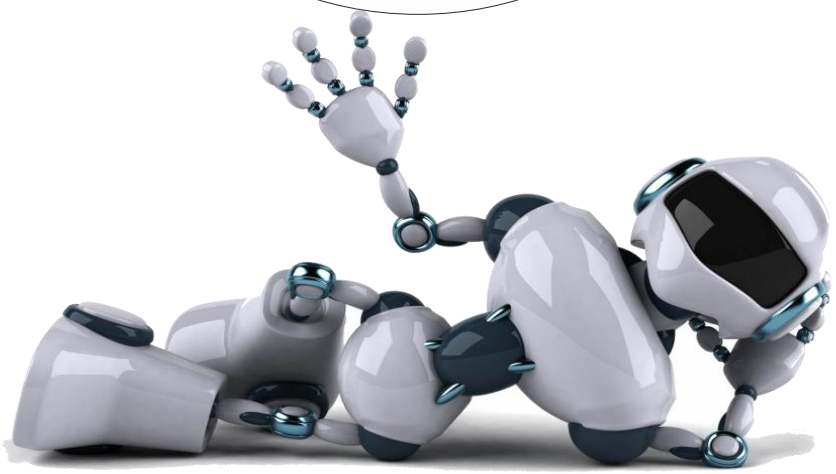
# Intelligence Artificielle générative et ÉDUcation :

## Enjeux, Défis et Perspectives de Recherche

IA-ÉDU 2025

@CORIA-TALN 2025

Ismail Badache



## Préface

Organisé par **Ismail Badache**, enseignant-chercheur à l'INSPÉ d'Aix-Marseille Université et rattaché au laboratoire LIS UMR 7020 CNRS. L'atelier IA-ÉDU (IA & ÉDUcation : Enjeux, Défis & Perspectives de Recherche) a pour objectif de rassembler chercheurs, enseignants (de l'école primaire à l'université), étudiants, professionnels et décideurs de diverses disciplines afin d'explorer les opportunités et les défis liés à l'intégration de l'intelligence artificielle générative (IAg) dans l'éducation et la formation. Les sujets abordés couvrent un large spectre interdisciplinaire et transdisciplinaire de recherches et d'applications, à la fois en informatique et en sciences de l'éducation et de la formation, incluant (sans s'y limiter) :

- L'agentivité numérique et l'appropriation critique des outils d'IA générative par les apprenants et les enseignants,
- L'analyse des données éducatives (trajectoires d'apprentissage, émotions, évaluations) pour une pédagogie plus adaptative,
- L'hybridation entre modèles génératifs et connaissances structurées (RAG, graphes de connaissances, bases vérifiées),
- L'évaluation et la génération automatique de contenus pédagogiques (questions, feedback, parcours personnalisés),
- L'inclusion et l'accessibilité grâce à l'IA (manuels scolaires adaptés, descriptions automatiques, outils pour les langues peu dotées),
- Les enjeux éthiques et cognitifs de l'IA conversationnelle dans l'éducation (autorité épistémique, biais, intégrité académique),
- Les approches hybrides combinant IA symbolique et connexionniste pour des technologies éducatives plus robustes,
- Les études de terrain sur l'intégration de l'IA en contexte scolaire (retours d'expérience, formations, perceptions des acteurs éducatifs).

L'atelier s'ouvrira par une conférence d'Abdellah Fourtassi, enseignant et chercheur à Aix-Marseille Université (LIS, UMR 7020, CNRS), intitulée : *Les grands modèles de langage peuvent-ils apprendre comme les enfants ?*

Cette conférence sera suivie d'une démonstration du robot éducatif Buddy, illustrant son intégration dans des dispositifs d'enseignement artistique. Ce moment permettra de découvrir concrètement comment la robotique sociale peut soutenir des pratiques pédagogiques innovantes. La matinée se conclura par une session de présentation de posters, mettant en lumière des travaux autour de l'IA, de l'apprentissage et de l'éducation.

L'après-midi débutera avec une intervention de Lucile Gelin, Data Scientist chez Lalilo by Renaissance Learning et chercheuse associée à l'IRIT UMR 5505 CNRS, sur le sujet : *Les utilisations de l'IA dans l'application Lalilo pour l'apprentissage de la lecture: objectifs, contraintes, évaluation.*

Elle sera suivie de plusieurs présentations de recherches interdisciplinaires, consacrées au développement de solutions numériques innovantes pour l'enseignement et l'apprentissage. Ces interventions mettront en valeur des approches combinant informatique, sciences de l'éducation et design pédagogique. Enfin, l'atelier se clôturera par une table ronde, réunissant l'ensemble des intervenants et les participants, dans un esprit de dialogue et de réflexion collective sur les enjeux actuels de l'intelligence artificielle dans l'éducation.

Nous exprimons notre profonde gratitude à l'INSPÉ d'Aix-Marseille ainsi qu'au laboratoire ADEF UR 4671 pour leur précieux soutien. Nous remercions également chaleureusement l'ensemble des personnes ayant contribué à la réussite de cet événement : les auteurs et autrices des soumissions, les membres du comité de programme, les membres du comité d'organisation, ainsi que l'équipe de coordination de CORIA-TALN 2025 pour leur appui logistique indispensable.

## Comité local d'organisation

- Ismail Badache, LIS, INSPÉ d'Aix-Marseille Université (**Président**)
- Caroline Vincent, LEST, INSPÉ d'Aix-Marseille Université
- Sarah Nouali, LIS, Aix-Marseille Université
- Abdelhak Dahmani, LIS, Aix-Marseille Université
- Mohsine Aabid, LIS, Aix-Marseille Université
- Paul Pouzergues, LPL, Aix-Marseille Université

## Comité de programme

- Ismail Badache, LIS, INSPÉ d'Aix-Marseille Université (**Président**)
- Pierre Bellet, DREAM-U, Aix-Marseille Université
- Patrice Bellot, LIS, Aix-Marseille Université
- Pascale Brandt-Pomares, ADEF, INSPÉ d'Aix-Marseille Université
- Abdellah Fourtassi, LIS, Aix-Marseille Université
- Gaël Guibon, LIPN, Université Sorbonne Paris Nord
- Maria Impedovo, ADEF, INSPÉ d'Aix-Marseille Université
- Patrice Laisney, ADEF, INSPÉ d'Aix-Marseille Université
- Jonathan Mirault, AMPIRIC, Aix-Marseille Université
- Éric Olivier, CIPE, Aix-Marseille Université
- Paul Pouzergues, LPL, Aix-Marseille Université
- Caroline Vincent, LEST, INSPÉ d'Aix-Marseille Université
- Aznam Yacoub, Université de Windsor, Canada

**Pierre Bellet**

**Patrice Bellot**

**Pascale Brandt-Pomares**

**Abdellah Fourtassi**

**Gaël Guibon**

**Maria Impedovo**

**Patrice Laisney**

**Jonathan Mirault**

**Éric Olivier**

**Paul Pouzergues**

**Caroline Vincent**

**Aznam Yacoub**

## Conférences invitées (Keynote 1)

**Abdellah Fourtassi, Maître de conférences, HDR en Informatique**
LIS UMR 7020 CNRS, Aix-Marseille Université, France.
Titre de la conférence : *Les grands modèles de langage peuvent-ils apprendre comme les enfants?*

**Résumé :** Les grands modèles de langage comme ChatGPT ont transformé l'intelligence artificielle, mais nécessitent des quantités massives de données — souvent des centaines de milliards de mots — pour atteindre un haut niveau de performance. En comparaison, les enfants humains acquièrent le langage à partir de seulement quelques millions de mots. Cet écart a suscité une nouvelle ligne de recherche visant à développer des modèles capables d'apprendre à partir d'un volume de données plus proche de celui des enfants. En explorant comment des systèmes peuvent apprendre le langage à partir de données limitées et naturelles, ces travaux offrent des perspectives utiles à la fois pour améliorer les modèles d'IA et pour mieux comprendre l'apprentissage humain. Cette recherche réunit des spécialistes de l'IA et des sciences cognitives, avec un double objectif : rendre les systèmes d'apprentissage plus efficaces en données, et mieux comprendre comment le langage se développe chez l'enfant dans des contextes d'apprentissage écologiques.

**Mots-clés :** IA, développement du langage, modélisation, apprentissage, cognition

**Abstract:** Large language models like ChatGPT have transformed the field of artificial intelligence, but they require vast amounts of data—often hundreds of billions of words—to reach high performance. In contrast, human children acquire language from just a few million words. This gap has motivated a new line of research focused on building language models that learn from human-scale input, closer to what children actually experience. By investigating how systems can learn language from limited, naturalistic data, this work offers insights that are valuable both for improving AI and for understanding human learning. It sits at the intersection of artificial intelligence and cognitive theories of language acquisition, bringing together researchers who aim to make learning systems more data-efficient and those who seek to model child language development in ecologically valid settings. This talk will explore the challenges and opportunities of this emerging research agenda and its implications.

**Keywords:** AI, language development, modeling, learning, cognition

## Conférences invitées (Keynote 2)

**Lucile GELIN, Senior Data Scientist - Lalilo & CHICA-AI ANR**
Renaissance Learning et IRIT UMR 5505 CNRS, Université de Toulouse, France
Titre de la conférence : *Les utilisations de l'IA dans l'application Lalilo pour l'apprentissage de la lecture: objectifs, contraintes, évaluation*

**Résumé :** La maîtrise de la lecture est une étape clé dans le développement de l'enfant pour qu'il devienne autonome. L'apprentissage n'est cependant pas facile, et les enseignant·es de primaire pourraient bénéficier de soutien dans cette tâche. Ce soutien peut être apporté par des applications éducationnelles, comme Lalilo, qui utilise l'intelligence artificielle pour proposer un outil de pratique qui s'adapte à chaque élève et renseigne l'enseignant·e pour faciliter la remédiation en classe. Cette présentation décrira les diverses utilisations de l'IA dans Lalilo, leurs objectifs pour les élèves et enseignant·es, et les contraintes pédagogiques et éthiques liées à leur intégration dans une application à destination de jeunes enfants. Nous discuterons des choix et processus mis en place pour assurer des performances adéquates et un impact bénéfique pour nos utilisateur·rices. Nous aborderons le sujet des relations entre recherche académique et R&D industrielle, et comment ces partenariats font avancer les technologies pour l'éducation.

**Mots-clés :** IA pour l'éducation, apprentissage adaptatif, reconnaissance de la parole

**Abstract:** Mastering reading is a key stage in a child's development towards independence. Learning is not easy, however, and primary school teachers would benefit from support in this task. This support can be provided by educational applications, such as Lalilo, which uses artificial intelligence to offer a practice tool that adapts to each student and informs the teacher to facilitate in-class remediation. This presentation will describe the various uses of AI in Lalilo, their objectives with students and teachers in mind, and the pedagogical and ethical constraints involved in integrating them into an application for young children. We'll discuss the choices and processes put in place to ensure adequate performance and a beneficial impact for our users. We'll also look at the relationship between academic research and industrial R&D, and how these partnerships are driving forward technologies for education.

**Keywords:** IA for education, adaptive learning, speech recognition

# Sommaire des articles de l'atelier IA-ÉDU

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Accessibilité visuelle et éducation inclusive : Étude préliminaire sur la génération de textes alternatifs

Elise Lincker[1] Elisabeth Olamisan[2] Theodora Pazakou[2]
Michèle Gouiffès[2] Camille Guinaudeau[2] Frédéric Dufaux[3]
(1) CNAM, Cedric, Paris, France (2) Université Paris-Saclay, CNRS, LISN, Orsay, France
(3) CentraleSupélec, Laboratoire des signaux et systèmes, Orsay, France
prenom.nom@lisn.fr

RÉSUMÉ

Tout contenu numérique devrait garantir l'accessibilité visuelle en incluant des textes alternatifs aux images. En l'absence de système et de métrique d'évaluation adaptés, nous présentons nos recherches préliminaires sur la génération et l'évaluation de textes alternatifs, d'abord dans un contexte générique. Dans une démarche d'inclusion scolaire, nous mettons en lumière les limites des systèmes existants et les contraintes à prendre en compte pour envisager un système applicable aux manuels scolaires.

ABSTRACT

**Visual Accessibility and Inclusive Education : A Preliminary Study on Alt-Text Generation**

All digital content should ensure visual accessibility by including alternative text for images. In the absence of suitable generation systems and evaluation metrics, we present our preliminary research on the generation and evaluation of alternative text, first in a generic context. In order to foster inclusive education, we highlight the limitations of existing systems and the challenges that must be addressed to develop a system applicable to school textbooks.

MOTS-CLÉS : Texte alternatif, Accessibilité visuelle, Éducation inclusive.

KEYWORDS: Alternative text, Alt text, Visual accessibility, Inclusive education.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

## 1 Introduction

L'accessibilité visuelle des contenus numériques constitue un enjeu majeur pour garantir l'inclusion des personnes déficientes visuelles (DV), aveugles et malvoyantes, dans l'accès à l'information. Il est donc essentiel de rendre les contenus visuels accessibles, soit en produisant des images tactiles pour les supports adaptés, soit, le plus souvent, en proposant un texte alternatif. Ce dernier est une description textuelle destinée à remplacer l'image pour les DV. Il peut être transmis avec l'ensemble du contenu textuel d'un document numérique à un dispositif d'accessibilité visuelle, tel qu'un lecteur d'écran ou une plage Braille. Des directives, comme celles de W3C pour l'accessibilité du Web [1], préconisent des descriptions concises, objectives et contextualisées qui mettent en évidence le contenu prédominant afin d'en faciliter la compréhension. Cependant, il n'existe pas de consensus universel sur les normes de l'accessibilité visuelle, et celles-ci sont peu appliquées.

1. https://www.w3.org/WAI/tutorials/images/

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 1–9.

Les manuels scolaires numériques restent majoritairement inaccessibles aux élèves DV, en raison d'une conception non inclusive et d'un manque de compatibilité avec les outils d'assistance (Castillan *et al.*, 2018). Bien que des recommandations existent pour intégrer l'accessibilité dès la conception [2], elles sont rarement appliquées, rendant nécessaires des adaptations a posteriori. Des organismes de transcription de documents pour un public DV transforment les contenus numériques en gros caractères, braille ou format audio comme DAISY. L'association autrichienne BookAccess propose notamment des directives spécifiques pour l'adaptation de manuels scolaires [3]. Le travail d'adaptation reste cependant largement manuel, ce qui le rend long, coûteux et dépendant d'une expertise humaine.

Cette étude préliminaire propose un cadre pour la génération et l'évaluation automatique de textes alternatifs, face à plusieurs défis : le contexte de l'image, les besoins spécifiques des DV, et l'absence de métrique d'évaluation appropriée. Nous identifions également les contraintes supplémentaires à prendre en compte en vue d'une application aux manuels scolaires.

## 2 Travaux connexes

Les études sur les préférences des DV mettent en évidence l'importance du contexte pour le texte alternatif. Il inclut notamment le scénario dans lequel apparaît l'image (l'objectif informationnel de l'utilisateur et la source de l'image) (Stangl *et al.*, 2021) ainsi que son contexte immédiat. Un même visuel peut ainsi nécessiter des descriptions différentes selon le scénario. Les DV préfèrent les descriptions contextualisées, jugées plus pertinentes et utiles lorsqu'elles répondent à leurs attentes en fonction du scénario (Gubbi Mohanbabu & Pavel, 2024; Kreiss *et al.*, 2022a).

Cependant, les modèles et métriques d'évaluation existants ne gèrent pas le contexte. Alors qu'une légende complète une image, le texte alternatif la remplace pour garantir l'accessibilité visuelle. Cette différence fondamentale empêche l'adoption directe des modèles de génération de légendes dans un cadre d'accessibilité, en raison du manque de prise en compte du contexte et de leur pré-entraînement sur des corpus non adaptés. À l'inverse, les modèles vision-langage (VLMs) tendent à produire des descriptions trop longues, comportant des spéculations et des hallucinations (Lincker *et al.*, 2025).

La génération de textes alternatifs reste un domaine encore peu exploré ou limité aux images scientifiques (Chintalapati *et al.*, 2022; Williams *et al.*, 2022; McCall & Chagnon, 2022). Pour des images de Twitter, Srivatsan *et al.* (2024) combinent CLIP (Radford *et al.*, 2021) et un réseau de mapping qui concatène les plongements visuels aux vecteurs contextuels, formant un préfixe pour la génération autorégressive d'une description. Une approche similaire a été adoptée par AutoAD (Han *et al.*, 2023) pour générer des audiodescriptions. Zur *et al.* (2024) affinent CLIP pour favoriser les descriptions aux légendes, mais sans amélioration de la pertinence visuelle, le contexte n'étant pas utilisé. Enfin, l'outil AltAuthor (Song *et al.*, 2025) aide les développeurs web à intégrer du texte alternatif. Il comprend la classification de l'image selon son rôle pour déterminer si elle nécessite un texte alternatif, la génération conforme aux normes le cas échéant, et une interface d'édition.

Toutefois, la transposition de ces approches au contexte scolaire reste difficile, en particulier en l'absence de jeux de données adaptés, et peu de travaux abordent la gestion des images pour l'éducation inclusive. Yadav *et al.* (2025) proposent une classification des illustrations dans les manuels d'Étude de la Langue selon leur rôle pédagogique : essentielles, informatives ou décoratives. Cette typologie oriente les adaptations en fonction du type de handicap et pourrait constituer un point de départ pour la génération de texte alternatif, à l'instar de l'approche d'AltAuthor.

2. https://www.firah.org/fr/access-man.html 3. https://www.bookaccess.at/

# 3 Méthodologie proposée

## 3.1 Données

Il n'existe actuellement aucun ensemble de données de contenus visuels pédagogiques adaptés aux DV. Des organismes proposent des adaptations de manuels scolaires à destination des élèves DV, mais ces contenus ne sont pas diffusés en raison de restrictions liées aux droits d'auteur. Dans d'autres domaines, des jeux de données ont été créés à partir de paires (image, texte) extraites automatiquement de pages web (Sharma *et al.*, 2018; Schuhmann *et al.*, 2022; Srivatsan *et al.*, 2024). Cependant, les descriptions sont souvent rédigées par des contributeurs non spécialistes de l'accessibilité visuelle, ce qui entraîne une qualité variable, rarement suffisante pour répondre aux besoins des DV. Par ailleurs, aucun corpus francophone de textes alternatifs n'est actuellement disponible.

Nos premières expérimentations s'appuieront sur les jeux de données en anglais Concadia (Kreiss *et al.*, 2022b) et AD2AT (Lincker *et al.*, 2025), qui se distinguent par la fiabilité de leurs textes alternatifs. Concadia contient 96 918 images issues de pages Wikipedia, accompagnées de leurs textes alternatifs, légendes et paragraphes contextuels. Les descriptions ont été filtrées pour garantir leur qualité. AD2AT est construit à partir d'audiodescriptions de films produites par des professionnels, dont la modalité a été transformée : de la vidéo avec audiodescription à l'image avec texte alternatif. Les audiodescriptions précédant l'image cible sont utilisées comme contexte. La partie AD2AT-MD contient 37 266 images extraites du dataset d'audiodescriptions MPII-MD (Rohrbach *et al.*, 2015). La deuxième section AD2AT-VIW est construite sur Visuals Into Words (Matamala & Villegas, 2016), un même film audiodécrit par dix descripteurs professionnels. Elle comprend 28 images, chacune associée à 1 à 10 textes alternatifs. Elle constitue une base de test pour évaluer la variabilité et la qualité des descriptions.

## 3.2 Génération

Les premières approches, reposant soit sur des modèles de génération de légendes, soit sur la formulation d'instructions à un VLM, ont montré leurs limites en contexte d'accessibilité, produisant souvent des descriptions génériques, trop longues ou spéculatives (Lincker *et al.*, 2025). Nous proposons d'affiner LLaVa (Liu *et al.*, 2023) par un nouveau réglage des instructions, en appliquant Low Rank Adaptation (LoRA) (Hu *et al.*, 2022) et Direct Preference Optimization (DPO) (Rafailov *et al.*, 2023). Cette approche permet d'éviter l'affinage complet ou l'apprentissage par renforcement, coûteux en ressources. DPO vise à aligner les LLMs sur les préférences humaines et nécessite que chaque exemple d'entraînement soit associé à une paire de références contrastées : l'une positive et l'autre négative.

Notre architecture, illustrée en Figure 1, repose sur un modèle LLaVa figé, dans lequel seuls les poids des matrices de faible rang A et B (matrices LoRA) sont ajustés pendant l'apprentissage. L'image et l'instruction (*prompt*), qui inclut le contexte, sont passés deux fois au modèle : avec une référence positive et avec une négative. La préférence du modèle est mesurée par la log-vraisemblance de chaque réponse : Score $= \log P_{\theta+\Delta_\theta}(\text{response} \mid \text{input})$, avec $\theta$ le modèle figé et $\Delta\theta$ les poids de LoRA. Le problème est posé comme une classification binaire, où la fonction de perte s'appuie sur la différence entre les deux scores : Loss $= -\log \sigma(\text{score\_ref\_pos} - \text{score\_ref\_neg})$ où $\sigma(x)$ est la fonction sigmoïde.

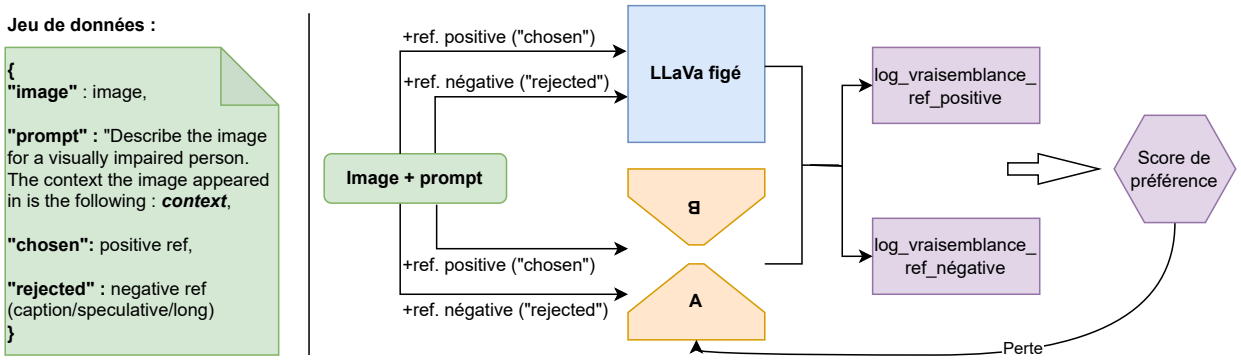



FIGURE 1 – Architecture proposée : Réglage des instructions de LLaVa avec LoRA et DPO

Pour l'entraînement, les références positives correspondent aux descriptions des jeux de données Concadia et AD2AT. Les légendes de Concadia servent de références négatives. En l'absence d'exemples négatifs pour AD2AT, nous en générons avec LLaVa. En ajustant uniquement les matrices LoRA, le modèle apprend à préférer des descriptions courtes, précises et adaptées au contexte, en rejetant les sorties trop longues ou spéculatives.

## 3.3 Evaluation

Les métriques classiques utilisées pour évaluer la génération de texte (BLEU (Papineni *et al.*, 2002), ROUGE (Lin & Hovy, 2003), METEOR (Banerjee & Lavie, 2005), CIDEr (Vedantam *et al.*, 2015), SPICE (Anderson *et al.*, 2016)) reposent sur la similarité avec des descriptions de référence produites par des humains. Cependant, elles ne tiennent compte ni du contexte d'apparition de l'image, ni des préférences des DV. De plus, une image dans un même contexte peut avoir plusieurs textes alternatifs acceptables, et ne doit pas nécessairement être évaluée en fonction d'une seule référence. Les résultats présentés dans (Kapur & Kreiss, 2024) soulignent la nécessité de développer une métrique sans référence qui prenne en compte les besoins spécifiques des DV. Les approches sans référence, telles que CLIPScore (Hessel *et al.*, 2021) et SPURTS (Feinglass & Yang, 2021), tentent de s'affranchir de la dépendance aux références. Toutefois, elles évaluent principalement la similarité brute entre l'image et le texte, sans prendre en compte la richesse de l'adéquation sémantique, la structure discursive, la cohérence logique ou les enjeux d'accessibilité (Hu *et al.*, 2023; Ahmadi & Agrawal, 2024). ContextRef (Kreiss *et al.*, 2024) constitue un banc d'essai pertinent, basé sur des paires image–texte contextualisées et des évaluations humaines, mais il se concentre sur une analyse de corrélation plutôt que sur une approche d'évaluation intégrée.

Pour pallier ces limites, nous proposons un cadre d'évaluation léger et interprétable. Le score final, **AltScore**, combine six dimensions : la *qualité d'ancrage visuel (QAV)*, la *cohérence discursive (CD)*, la *fluidité et cohérence linguistique (FCL)*, l'*imageabilité (IMG)*, la *pertinence contextuelle (PC)* et la *plausibilité des concepts de sens communs (PCC)*.

$$\text{AltScore} = \lambda_1 \cdot \text{QAV} + \lambda_2 \cdot \text{CD} + \lambda_3 \cdot \text{FCL} + \lambda_4 \cdot \text{IMG} + \lambda_5 \cdot \text{PC} + \lambda_6 \cdot \text{PCC}$$

**Qualité d'ancrage visuel**. Nous extrayons des masques de segmentation au niveau des objets à l'aide d'un modèle de segmentation tel qu'EfficientSAM (Xiong *et al.*, 2024). Chaque région segmentée ainsi que l'image entière sont encodées indépendamment à l'aide d'un modèle vision-langage compact (par exemple, SmolVLM (Marafioti *et al.*, 2025)). Les rôles sémantiques sont extraits du texte

alternatif à l'aide de techniques standards d'étiquetage des rôles sémantiques (Chen *et al.*, 2025). Chaque représentation vectorielle des rôles sémantiques est comparée aux représentations enrichies des régions visuelles, en calculant la similarité maximale (MaxSim) pour chaque rôle. Le *contexte local* fait référence aux régions segmentées spécifiques, tandis que le *contexte global* correspond à l'image dans son ensemble. La fusion des deux permet une évaluation fine de l'alignement entre le texte et le contenu visuel.

**Cohérence discursive**. Lorsque le texte alternatif comporte plusieurs phrases, nous en évaluons la cohérence interne. Nous vérifions que les entités (acteurs, objets) sont référencées de manière cohérente, qu'aucune contradiction n'est introduite, et que les rôles sémantiques à travers les phrases forment une description cohérente de la scène. La résolution des coréférences et l'analyse des rôles sémantiques sont utilisées pour soutenir cette étape de validation.

**Évaluation de la fluidité et de la cohérence linguistique**. La qualité linguistique de surface est évaluée selon une dimension inspirée de GRUEN (Zhu & Bhat, 2020), combinant trois sous-scores : la grammaticalité, la non-redondance et la cohérence thématique.

**Imageabilité**. Nous estimons la capacité évocatrice du texte alternatif en suivant une version adaptée de la méthode Tell as You Imagine (Umemura *et al.*, 2021). Des scores d'imageabilité issus de lexiques psycholinguistiques sont attribués aux mots clés et agrégés le long de l'arbre syntaxique. La méthode est adaptée pour pénaliser l'abstraction et encourager une visualisation concrète.

**Pertinence contextuelle**. Nous encodons le contexte textuel immédiat de l'image (via MiniLM (Wang *et al.*, 2020)) afin de pondérer la pertinence des régions visuelles : un texte proche et pertinent les renforce, tandis qu'un contexte éloigné ou non pertinent a peu d'effet.

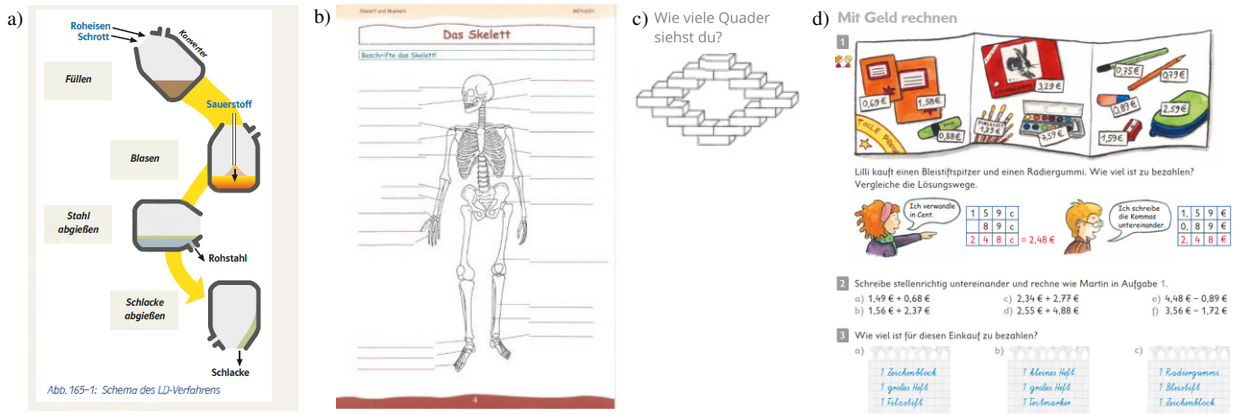
**Validation des concepts de sens commun**. Nous ajoutons une validation optionnelle de la plausibilité des rôles sémantiques. Les rôles sémantiques (acteurs, actions, objets) extraits du texte alternatif sont encodés sous forme de plongements de phrases, puis comparés à des représentations pré-encodées de scénarios plausibles issus du sens commun. Les combinaisons sémantiques présentant une faible similarité avec des événements typiques sont signalées comme potentiellement invraisemblables (Kapur & Kreiss, 2024).

Nous prévoyons d'évaluer dans quelle mesure **AltScore** est aligné avec les préférences des utilisateurs aveugles ou malvoyants, en corrélant ses résultats avec des jugements humains.

## 3.4 Application aux manuels scolaires

Les illustrations dans les manuels scolaires jouent un rôle crucial dans l'apprentissage. Lorsqu'une activité contient une image, l'adaptation diffère selon le rôle de celle-ci. L'élève doit pouvoir résoudre l'activité, grâce à une description de l'image ou un ajustement de la tâche. La figure 2 présente des extraits de manuels avec une adaptation des images pour les élèves DV. L'image a) est un schéma informatif essentiel à la compréhension d'une leçon ; elle doit être décrite en détail. L'image b) est un schéma à compléter par l'élève : sa description seule n'ayant aucun sens pour un élève DV, la tâche a été reformulée. L'image c) ne peut être décrite sans dévoiler la solution ; on indique la présence d'un graphique, et, si possible, un support tactile est proposé. Enfin, l'activité d) contient plusieurs images. La première fournit des informations essentielles à la réalisation des exercices et est remplacée par un texte alternatif clair et concis. Les énoncés des exercices à la suite sont également complétés pour inclure les informations visuelles, afin de ne pas alourdir la charge cognitive de l'élève.

Dans un objectif d'automatisation, il est nécessaire d'identifier les activités associées aux images, puis de filtrer celles-ci en fonction de leur nature et de leur rôle, en s'appuyant par exemple sur la classification proposée par Yadav *et al.* (2025). Les images décoratives, redondantes avec le texte, ou




Adaptations traduites de l'allemand vers l'anglais :

a) [[Graphic: Fig. 165-1: Diagram of the basic oxygen process:
Converter with 2 adjacent openings in different rotated positions
1. filling - of pig iron and scrap through the large opening (pointing diagonally upwards)
2. blowing - of oxygen through the large opening (pointing upwards)
3. tapping steel - pouring out raw steel from the small opening (pointing diagonally downwards)
4. tapping slag - from the large opening (now pointing downwards)]]

b) [[Which bones belong:
-) to the skull: []
-) to the trunk: []
-) to the limbs: []]]

c) [[Graphic: Geometric figure]]
[[Task solvable with tactile material]]

d) [[Graphic: Prices
small notebook: €0.69
large notebook: €1.58
...
pencil sharpener: €1.59]]
...
How much must be paid for this purchase?
a) 1 drawing pad (€3.29), 1 large notebook (€1.58), 1 felt-tip pen (€0.75)
[]

FIGURE 2 – Exemples d'images et activités de manuels scolaires et propositions d'adaptations par BookAccess. Source : https://www.bookaccess.at/

purement illustratives sans lien avec l'activité seront omises. En fonction de l'objectif pédagogique, il s'agit ensuite d'inclure une description de l'image et/ou de modifier la tâche pour la rendre accessible à un élève DV, si cela est possible. Nous envisageons un autre système de génération entièrement adapté au contexte scolaire. Les VLMs génériques bruts s'avèrent insuffisants dans ce cadre, d'autant plus qu'il n'existe pas de données d'entraînement dans un cadre scolaire. Par exemple, pour l'image du squelette (Figure 2 b)), même en contextualisant l'image dans l'instruction, LLaVa produit une description non adaptée comme « A skeleton is labeled with various body parts such as the skull, ribcage, pelvis, and legs. The skeleton stands on one leg. The image is in black and white. » Un tel système devrait prendre en compte les besoins des élèves DV et les objectifs pédagogiques. Il serait paramétrable à la fois en niveau de détail (selon la nature et le rôle de l'image) et en vocabulaire (adapté au niveau scolaire de l'enfant).

# 4 Conclusion

Dans une démarche inclusive, ce travail exploratoire propose un cadre pour la génération de textes alternatifs aux images tenant compte du contexte et des besoins des DV, en exploitant l'IA générative et en contournant ses limites. En l'absence de métrique d'évaluation adaptée, nous introduisons un score sans référence, avec pour perspective de l'aligner aux préférences des DV. Enfin, nous mettons en évidence la complexité de l'adaptation des contenus visuels dans les documents pédagogiques, en fonction de leur rôle et en raison de l'absence de jeux de données adaptés. Nos travaux futurs visent à appliquer et optimiser les méthodes proposées dans un contexte générique, puis à développer un système plus complexe spécifique aux activités scolaires.

# Remerciements

Ce travail a été financé par le projet ANR-21-CE38-0014, l'institut DATAIA et le LISN.


# Références


AHMADI S. & AGRAWAL A. (2024). An examination of the robustness of reference-free image captioning evaluation metrics. In *Findings of the Association for Computational Linguistics : EACL 2024*, p. 196–208.

ANDERSON P., FERNANDO B., JOHNSON M. & GOULD S. (2016). SPICE : Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, p. 382–398.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

CASTILLAN L., LEMARIÉ J. & MOJAHID M. (2018). Numérique, handicap visuel et accessibilité des apprentissages. contenus pédagogiques numériques : quelle accessibilité pour les élèves présentant une déficience visuelle ? *Éducation & Formation*, p. 90–102.

CHEN H., ZHANG M., LI J., ZHANG M., ØVRELID L., HAJIČ J. & FEI H. (2025). Semantic role labeling : A systematical survey. *arXiv preprint arXiv :2502.08660*.

CHINTALAPATI S. S., BRAGG J. & WANG L. L. (2022). A dataset of alt texts from HCI publications : Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, p. 1–12.

FEINGLASS J. & YANG Y. (2021). SMURF : SeMantic and linguistic UndeRstanding Fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 2250–2260.

GUBBI MOHANBABU A. & PAVEL A. (2024). Context-aware image descriptions for web accessibility. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, p. 1–17.

HAN T., BAIN M., NAGRANI A., VAROL G., XIE W. & ZISSERMAN A. (2023). AutoAD : Movie Description in Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 18930–18940.

HESSEL J., HOLTZMAN A., FORBES M., LE BRAS R. & CHOI Y. (2021). CLIPScore : A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7514–7528.

HU A., CHEN S., ZHANG L. & JIN Q. (2023). InfoMetIC : An informative metric for reference-free image caption evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3171–3185.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). LoRA : Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

KAPUR R. & KREISS E. (2024). Reference-based metrics are biased against blind and low-vision users' image description preferences. In *Proceedings of the Third Workshop on NLP for Positive Impact*, p. 308–314.

KREISS E., BENNETT C., HOOSHMAND S., ZELIKMAN E., MORRIS M. R. & POTTS C. (2022a). Context matters for image descriptions for accessibility : Challenges for referenceless evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 4685–4697.

KREISS E., FANG F., GOODMAN N. & POTTS C. (2022b). Concadia : Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 4667–4684.

KREISS E., ZELIKMAN E., POTTS C. & HABER N. (2024). ContextRef : Evaluating Referenceless Metrics For Image Description Generation. In *The Twelfth International Conference on Learning Representations*.

LIN C.-Y. & HOVY E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, p. 150–157.

LINCKER E., GUINAUDEAU C. & SATOH S. (2025). AD2AT : Audio description to alternative text, a dataset of alternative text from movies. In *Proceedings of the 31st International Conference on Multimedia Modeling*, p. 58–71.

LIU H., LI C., WU Q. & LEE Y. J. (2023). Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, **36**, 34892–34916.

MARAFIOTI A., ZOHAR O., FARRÉ M., NOYAN M., BAKOUCH E., CUENCA P., ZAKKA C., ALLAL L. B., LOZHKOV A., TAZI N. *et al.* (2025). Smolvlm : Redefining small and efficient multimodal models. *arXiv preprint arXiv :2504.05299*.

MATAMALA A. & VILLEGAS M. (2016). Building an audio description multilingual multimodal corpus : the VIW project. *Multimodal Corpora : Computer vision and language processing*, (11).

MCCALL K. & CHAGNON B. (2022). Rethinking alt text to improve its effectiveness. In *International Conference on Computers Helping People with Special Needs*, p. 26–33.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G. & SUTSKEVER I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, p. 8748–8763.

RAFAILOV R., SHARMA A., MITCHELL E., MANNING C. D., ERMON S. & FINN C. (2023). Direct Preference Optimization : Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, **36**, 53728–53741.

ROHRBACH A., ROHRBACH M., TANDON N. & SCHIELE B. (2015). A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 3202–3212.

SCHUHMANN C., BEAUMONT R., VENCU R., GORDON C., WIGHTMAN R., CHERTI M., COOMBES T., KATTA A., MULLIS C., WORTSMAN M., SCHRAMOWSKI P., KUNDURTHY S., CROWSON K., SCHMIDT L., KACZMARCZYK R. & JITSEV J. (2022). Laion-5b : An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, **35**, 25278–25294.

SHARMA P., DING N., GOODMAN S. & SORICUT R. (2018). Conceptual captions : A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2556–2565.

SONG H., SHIN M., KIM Y., JANG K., CHOI J., JUNG H. & SUH B. (2025). Altauthor : A context-aware alt text authoring tool with image classification and lmm-powered accessibility compliance. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, p. 124–128.

SRIVATSAN N., SAMANIEGO S., FLOREZ O. & BERG-KIRKPATRICK T. (2024). Alt-text with context : Improving accessibility for images on twitter. In *The Twelfth International Conference on Learning Representations*.

STANGL A., VERMA N., FLEISCHMANN K. R., MORRIS M. R. & GURARI D. (2021). Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd international ACM SIGACCESS conference on computers and accessibility*, p. 1–15.

UMEMURA K., KASTNER M. A., IDE I., KAWANISHI Y., HIRAYAMA T., DOMAN K., DEGUCHI D. & MURASE H. (2021). Tell as you imagine : Sentence imageability-aware image captioning. In *Proceedings of the 27th International Conference on MultiMedia Modeling, Part II*.

VEDANTAM R., LAWRENCE ZITNICK C. & PARIKH D. (2015). CIDEr : Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4566–4575.

WANG W., WEI F., DONG L., BAO H., YANG N. & ZHOU M. (2020). Minilm : Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, **33**, 5776–5788.

WILLIAMS C., DE GREEF L., HARRIS III E., FINDLATER L., PAVEL A. & BENNETT C. (2022). Toward supporting quality alt text in computing publications. In *Proceedings of the 19th International Web for All Conference*, p. 1–12.

XIONG Y., VARADARAJAN B., WU L., XIANG X., XIAO F., ZHU C., DAI X., WANG D., SUN F., IANDOLA F., KRISHNAMOORTHI R. & CHANDRA V. (2024). EfficientSAM : Leveraged Masked Image Pretraining for Efficient Segment Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 16111–16121.

YADAV S., LINCKER E., HURON C., MARTIN S., GUINAUDEAU C., SATOH S. & SHUKLA J. (2025). Towards inclusive education : Multimodal classification of textbook images for accessibility. In *Proceedings of the 31st International Conference on Multimedia Modeling*, p. 212–225.

ZHU W. & BHAT S. (2020). GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 94–108.

ZUR A., KREISS E., D'OOSTERLINCK K., POTTS C. & GEIGER A. (2024). Updating clip to prefer descriptions over captions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 20178–20187.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Annotation de résumés oraux d'élèves de primaire pour l'analyse automatique des capacités de compréhension de la lecture

Etienne Labbé[2] Brice Brossette[3] Nathalie Camelin[4] Tiphaine Caudrelier[4] Eddy Cavalli[4] Isabelle Ferrané[2] Barbara Lutz[1] Véronique Moriceau[2] Thomas Pellegrini[2] Julien Pinquier[2] Cantin Prat[4] Lucile Gelin[1, 2]
(1) Lalilo by Renaissance Learning, France
(2) IRIT, Université Paul Sabatier, CNRS, Toulouse, France
(3) Laboratoire d'Etude des Mécanismes Cognitifs (EMC), Université Lumière Lyon 2, Lyon, France
(4) Laboratoire d'Informatique d'Avignon (LIA), Avignon, France
etienne.labbe@irit.fr, lucile.gelin@renaissance.com

## RÉSUMÉ

Le projet CHICA-AI vise à construire une activité assistée par ordinateur pour l'entraînement des compétences de compréhension de la lecture des élèves de primaire. Cette activité consiste à demander à l'élève de résumer à l'oral un texte narratif, afin d'identifier ses difficultés de compréhension et fournir un retour personnalisé à l'élève et à son enseignant. Pour cela, nous mettrons en place un système automatique d'analyse fine des résumés oraux, capable d'extraire les informations pertinentes et de les combiner pour remplir une grille de critères pédagogiques et psycho-cognitifs. Nous présentons ici les défis du projet, ainsi que les premiers travaux réalisés : création de l'activité dans la plateforme Lalilo et du contenu pédagogique, collecte d'enregistrements audios, construction du protocole d'annotation. Nous présentons enfin les analyses préliminaires faites sur les premières annotations, qui serviront à l'entraînement et l'évaluation de notre système automatique.

## ABSTRACT

**Annotation of oral summaries from primary school students for automatic analysis of reading comprehension skills**

The CHICA-AI project aims to build a computer-assisted learning activity for training the reading comprehension skills of primary school pupils. This activity involves asking students to orally summarize a narrative text, in order to identify their comprehension difficulties and provide personalized feedback to the pupil and their teacher. To do this, we'll be implementing an automatic system for fine-grained analysis of oral summaries, capable of extracting relevant information and combining it to fill out a grid of pedagogical and psycho-cognitive criteria. We present here the challenges of the project, as well as the first tasks carried out : creation of the activity in the Lalilo platform and of the pedagogical content, audio recordings collection, construction of the annotation protocol. Finally, we present the preliminary analyses executed on the first annotations, which will be used to train and evaluate our automatic system.

MOTS-CLÉS : Traitement automatique de la parole et du language, apprentissage de la lecture, IA pour l'éducation.

KEYWORDS: Natural language and speech processing, reading learning, AI for education.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 10–16.

# 1 Introduction

La maîtrise de la lecture est essentielle à l'autonomie de l'enfant, mais son apprentissage est un véritable défi. En France, les compétences en lecture et en compréhension des élèves de CM1 ont baissé depuis le début des années 2000. L'étude PIRLS de 2016 révèle que 40,5% des jeunes de 15 ans ne maîtrisent pas la lecture et que 21,5% d'entre eux rencontrent de sérieuses difficultés (Colmant & Le Cam, 2017). L'enquête PISA de 2015 a mis en évidence des écarts de performance importants entre les élèves les plus performants et les moins performants(OCDE, 2016). L'enquête PISA de 2018 a montré que les élèves socio-économiquement favorisés dépassent leurs pairs défavorisés d'environ quatre années de scolarité(OCDE, 2019). Enfin, un rapport de l'Observatoire national de la lecture de 2007 [1] a établi un lien entre les problèmes de lecture et les échecs dans diverses matières. Une solution à ces problèmes pourrait résider dans un apprentissage assisté par ordinateur (*Computer-Assisted Learning*, CAL) efficace pour aider les enseignants et les élèves.

L'entreprise Lalilo développe un tel système CAL : un assistant pédagogique basé sur l'intelligence artificielle (IA) pour différencier l'apprentissage de la lecture. Ce système offre une grande variété de tâches associatives pour couvrir les deux aspects clés de la lecture : la reconnaissance des mots et la compréhension du langage. Ce dernier aspect est toutefois plus difficile à traiter par les systèmes CAL en raison de la complexité des processus impliqués dans la compréhension. Les exercices à réponse ouverte, par opposition aux QCM traditionnels, peuvent aider les étudiants à s'exercer à des aspects plus complexes de la compréhension. Dans ce contexte, Lalilo souhaite proposer une nouvelle intervention CAL avec une activité de résumé oral d'un texte, des retours personnalisés pour les élèves et des métriques informatives pour les enseignants.

L'objectif du projet CHICA-AI est de développer le système automatique CAL qui analysera et évaluera les résumés oraux des élèves. Nous avons choisi d'évaluer les résumés non seulement à un niveau global, mais surtout à un niveau plus fin, comme le font les psychologues et les enseignants en classe. Dans cet objectif, nous avons conçu une grille d'évaluation basée sur les sciences psycho-cognitives et la pédagogie. Le système CAL visera à extraire les informations pertinentes du résumé oral de l'élève et à donner une note pour chaque critère de la grille. Les technologies suivantes seront utilisées pour extraire les informations : reconnaissance automatique de la parole (RAP), compréhension du langage parlé (CLP), traitement automatique du langage naturel (TALN).

Pour mener à bien cet objectif, nous devrons surmonter deux défis majeurs.

**Défi n°1 : Adaptation de technologies à un cas d'utilisation complexe de la vie réelle**

- Parole d'enfant : la performance des modèles de RAP sur la parole d'enfant est loin de la performance sur les adultes, et il n'y a pas de systèmes de CLP existants qui sont adaptés à la parole d'enfant ;
- Parole spontanée avec une charge mentale élevée : la tâche de résumé oral nécessite des mécanismes cognitifs complexes, ce qui introduit à des événements de parole spontanée dans le discours de l'enfant, tels que les disfluences, répétitions, auto-corrections, hésitations, etc. La présence de ces événements augmente la difficulté des tâches automatiques ;
- Utilisation en classe : la plateforme Lalilo est conçue pour une utilisation à l'école. Cela implique que les enregistrements contiennent une grande variété de bruits de classe, qui peuvent nuire à la performance des systèmes de RAP.

1. «La lecture au début du collège», téléchargeable : http://onl.inrp.fr/ONL/publications/publi2007/

**Défi n°2 : Construction d'une activité complexe d'entraînement de la compréhension**

Cette intervention CAL est inédite et doit être créée de toute pièce. Les aspects suivants devront être pris en compte pour garantir l'efficacité de l'intervention sur les progrès en lecture de l'élève :

1. Expérience psycho-cognitive et pédagogique
   (a) Evaluation du résumé : Comment évaluer un résumé oral ? Quels sont les critères importants et ceux qui le sont moins ? Comment déterminer si un critère a été satisfait ?
   (b) Remédiation à la suite de l'activité : Comment détecter les difficultés des élèves ? Lorsque ces difficultés sont détectées, de quoi l'élève a-t-il besoin pour les surmonter ?
2. Défis techniques
   (a) Annotation des données : Comment annoter les données pour entraîner les modèles à remplir la grille d'évaluation ? L'évaluation est très subjective (même chez les enseignants expérimentés), comment assurer la qualité et la représentativité des annotations ?
   (b) Remplir automatiquement la grille d'évaluation : Comment combiner les informations extraites automatiquement pour répondre à chaque critère de la grille d'évaluation ? Comment mesurer la performance du système ?

Nous présentons dans cet article le travail effectué depuis le début du projet CHICA-AI. La première étape (section 2), a été de créer l'activité de résumé oral dans la plateforme Lalilo, ce qui nous a permis de récolter des enregistrements audios de résumés d'élèves. La seconde étape (section 3) a été de mettre en place une tâche d'annotation et un protocole rigoureux pour assurer la bonne qualité et la pertinence des annotations pour l'entraînement et l'évaluation des systèmes automatiques. Nous présentons en section 4 les analyses préliminaires de notre processus d'annotation.

# 2 L'activité de résumé oral

## 2.1 Déroulement de l'activité

L'activité de résumé oral est déployée dans la plateforme Lalilo depuis mai 2024. Les élèves suivent une progression pédagogique définie par l'équipe pédagogique de Lalilo, et atteignent l'activité de résumé oral au niveau CM1-CM2.

L'activité de résumé oral est précédée par plusieurs leçons pour apprendre aux élèves à extraire les informations pertinentes d'un texte et à construire un résumé à partir de ces informations. Une fois ces leçons terminées, les élèves ont accès à l'activité, qu'ils peuvent effectuer (pour l'instant) 5 fois, avec 5 textes différents. Les textes sont des textes narratifs créés par l'équipe pédagogique de Lalilo avec une difficulté croissante. L'activité est composée de plusieurs phases :

1. Lecture du texte par l'élève ;
2. Réponse à des questions de compréhension sur le texte pour l'aider à extraire les informations pertinentes. Plus l'élève est avancé dans sa progression, moins il y a de questions ;
3. Enregistrement du résumé oral ;
4. Validation par l'enfant de son résumé.

Au terme du projet, l'activité comportera une cinquième étape, durant laquelle l'enfant recevra un retour personnalisé à l'aide du système d'analyse automatique. L'objectif est de l'aider à mettre en place des stratégies pour progresser en compréhension de la lecture.

## 2.2 Grille d'évaluation des résumés

Nous avons construit une grille à partir de critères pédagogiques et psycho-cognitifs pour évaluer de façon fine les résumés des élèves et trouver les meilleures stratégies de remédiation. Nous nous sommes basés sur la grille de (Casazza, 1993), que nous avons adaptée pour qu'elle corresponde à des textes narratifs niveau primaire et à du travail de compréhension de la lecture (Cèbe *et al.*, 2004).

Les critères sont divisés en deux catégories : stratégie de compréhension et mise en forme du résumé. La première catégorie contient des critères sur le contenu du résumé : les personnages principaux et secondaires, les lieux de l'histoire, les idées principales à retrouver et leur chronologie dans le récit de l'élève. La seconde catégorie permet d'évaluer la qualité de la forme du résumé : reformulation des idées, présence de connecteurs logiques et temporels, mise en oeuvre des compétences syntaxiques et sémantiques, etc.

# 3 Construction des annotations

## 3.1 Tâche d'annotation

La tâche d'annotation est réalisée sur une application développée par nos soins et fondée sur la bibliothèque python Streamlit[2]. Cette application permet de facilement traiter les enregistrements les uns après les autres et enregistre les annotations sous format json. Le processus d'annotation est composé de plusieurs phases.

La première phase consiste à écarter les enregistrements non annotables, qui ne seront pas utilisés dans le projet. En effet, les enfants étant en autonomie sur la plateforme Lalilo, il arrive souvent que les enregistrements reçus ne présentent pas de contenu analysable. Nous rejettons les enregistrements pour les raisons suivantes :

- NO_VOICE : L'élève ne parle pas
- ADULT : On entend seulement un adulte qui parle (souvent l'enseignant)
- OUT_OF_EXERCISE : L'élève parle mais ne fait pas l'exercice
- TOO_NOISY : L'enfant parle mais il y a trop de bruit pour comprendre ce qui est dit
- NOT_INTELLIGIBLE : L'enfant parle de façon non intelligible, on ne le comprend pas

La seconde phase, si l'enregistrement n'est pas rejeté, consiste à transcrire le résumé le plus fidèlement possible. Cette transcription sera utilisée pour entraîner les futurs systèmes de RAP et CLP, et servir de base textuelle pour les systèmes de TALN. Nous utilisons le système Whisper-large-v3[3] pour établir une transcription automatique à corriger. Des symboles sont ajoutés manuellement pour transcrire les pauses, hésitations, sons non verbaux et mots inintelligibles.

La troisième phase comporte les critères de la catégorie "stratégie de compréhension" de la grille. Les personnages, lieux et idées font l'objet de deux annotations : un score d'identification (1-4) et un repérage d'entités nommées (consistant à surligner des mots dans la transcription du résumé). La chronologie est évaluée avec un système de classement des idées principales identifiées par l'élève. L'annotation correspond à la chronologie telle qu'elle est dans l'esprit de l'élève, et non telle que présentée dans le résumé, c'est à dire en prenant compte de l'utilisation de connecteurs temporels (par exemple "avant ça", "plus tôt dans l'histoire"...). Cette phase se termine par un score global, entre

2. https://streamlit.io/
3. https://huggingface.co/openai/whisper-large-v3

0 et 10, du contenu du résumé de l'élève, représentant sa compréhension de l'histoire.

La quatrième phase correspond aux critères de la catégorie "mise en forme du résumé" qui ne sont pas annotables automatiquement. Nous cherchons à détecter la présence d'avis personnels de l'enfant sur l'histoire (à proscrire dans un résumé) ou de parole hors de l'exercice (signe d'une déconcentration de l'élève). Nous demandons également à l'annotatrice de donner, entre 1 et 5, un score global de mise en forme du résumé et une score global d'expression orale.

## 3.2 Protocole d'annotation

En premier lieu, nous avons mené en 2024 une étude pilote à plus petite échelle sur les données collectées au cours des premiers mois sur la plateforme Lalilo. Dans cette étude préliminaire, une première version du guide d'annotation et de l'interface d'annotation a été conçue pour répondre à la tâche décrite ci-dessus. L'étude a été réalisée avec des étudiantes orthophonistes et a permis d'obtenir divers retours pour améliorer et corriger certains points de l'interface d'annotation et du guide. Après amélioration du protocole, une première réelle salve d'annotation a été réalisée en 2025 avec six autres étudiantes orthophonistes. Pour les former à cette tâche d'annotation très complexe, nous avons divisé le processus d'annotation en différentes étapes :

1. Formation initiale : présentation du protocole d'annotation (guide, interface) ;
2. Lot de formation (n°1) : Annotation d'un lot de 15 enregistrements choisis pour couvrir une diversité de cas (un exemple simple, moyen et complexe pour chaque histoire), chaque enregistrement étant annoté par toutes les annotatrices ;
3. Mesure de la qualité des annotations sur le lot 1 : Calcul de différentes mesures inter-annotateurs pour chaque domaine. L'objectif est de s'assurer que chaque annotatrice a bien compris les lignes directrices et fournit des annotations cohérentes et correctes.
4. Formation complémentaire et correction collaborative du lot 1
5. Lot de validation (n°2) : Annotation d'un lot de 45 audios répartis par paires d'annotatrices ;
6. Mesure de la qualité des annotations sur le lot 2 : Si les accords inter-annotateurs sont en dessous de nos seuils prédéfinis pour la validation, retour à l'étape 4. Dans le cas contraire, nous procédons à l'annotation du lot final.
7. Lot final : Le reste des données est réparti entre les annotatrices, en conservant 20% d'enregistrements en commun pour calculer des accords inter-annotatrices.

Pour estimer la qualité des annotations, nous avons utilisé plusieurs métriques différente en fonction de la nature du champ annoté. Les scores globaux sont évalués par le coefficient de corrélation intraclasse (*Intraclass Correlation Coefficient*, ICC) (Koch, 2006), et plus précisément l'ICC3 (modèle à deux facteur mixte). Les séquences (transcriptions et chronologies) sont comparées par le ratio de la distance de Levenshtein (au niveau du mot pour les transcriptions et au niveau de l'index de l'idée pour les chronologies). Les autres critères catégoriques sont évalués par le Kappa de Cohen.

# 4 Résultats

Le jeu de données obtenu contient au total 2085 annotations, dont 1065 ont été rejetés par les annotatrices. La majorité des rejets concernent soit des fichiers audio ne contenant aucune parole (481 annotations), soit des fichiers audio ne contenant que des paroles hors de l'exercice (364 annotations). Il en résulte 1020 annotations de 875 fichiers audio différents.

Les résultats des accords inter-annotateurs sont donnés dans la table 1. Les accords sur la transcription sont restés supérieurs au seuil de 0,85 pour tous les lots. Dans le lot 1, 2 et final, les accords sur les scores de compréhension (compréhension globale, personnages principaux, etc.), étaient au-dessus du seuil attendus pour l'ICC3 (0,6), et ont même dépassé 0,9, indiquant que les annotatrices ont bien compris ces critères. Pour la chronologie, les ratios sont plus bas, dû au fait qu'une idée non-reconnue par une annotatrice mais reconnue par une autre pour un même fichier fait légèrement baisser le score. Si on ne considère que les idées en commun, l'accords est bien plus élevé (>0,97). Les scores de mise en forme et d'expression globaux sont plus difficiles à annoter et plus variables. Après le lot 2, ils étaient au dessus du seuil acceptable (0,6), mais ont chuté à 0,50 et 0,55 pour le lot final. En dépit de cela, les accords peuvent être considérés comme globalement satisfaisants en raison de la complexité de l'annotation mise en œuvre, et ils suggèrent que le protocole pourra être appliqué à nouveau pour annoter les résumés oraux futurs de la plateforme.

TABLE 1 – Résultats des accords inter-annotateurs sur les lots d'entrainement 1 et 2 ainsi que sur les données communes du lot final.

| **Champ d'annotation** | **Métrique** | **Lot 1** | **Lot 2** | **Lot final** |
|---|---|---|---|---|
| Transcription | Levenshtein | 0,9102 | 0,8559 | 0,8781 |
| Compréhension globale | ICC3 | 0,6510 | 0,9089 | 0,8141 |
| Personnages principaux | ICC3 | 0,6775 | 0,9282 | 0,9306 |
| Idées principales | ICC3 | 0,7303 | 0,8092 | 0,8859 |
| Lieux | ICC3 | 0,9666 | 0,8922 | 0,8836 |
| Personnages secondaires | ICC3 | 0,7816 | 0,9666 | 0,8633 |
| Chronologie des idées | Levenshtein | 0,7776 | 0,8135 | 0,8551 |
| Mise en forme globale | ICC3 | 0,3173 | 0,7517 | 0,5092 |
| Expression globale | ICC3 | 0,6945 | 0,6708 | 0,5594 |

# 5 Conclusion

Nous avons présenté dans ce travail les objectifs et premiers résultats du projet CHICA-AI, visant à fournir un entraînement à la compréhension de la lecture aux élèves de primaire à travers une activité de résumé oral avec analyse automatique et retour personnalisé. Les premiers travaux effectués ont permis de récolter des enregistrements audios qui seront utilisés pour entraîner le système d'analyse automatique des résumés, ainsi que de les annoter pour les différentes tâches envisagées (RAP, CLP, TALN). Nos analyses montrent que notre protocole d'annotation, très sophistiqué en raison de la complexité de la tâche, permet d'obtenir des annotations de qualité satisfaisante pour la plupart des critères d'évaluation d'un résumé. Les chercheurs et chercheuses impliquées dans le projet pourront ainsi utiliser ces annotations pour entraîner et évaluer leurs systèmes, et ainsi construire une activité bénéfique dont l'usage entraîne une amélioration dans les capacités de compréhension des élèves.

# Remerciements

Les auteurs remercient l'Agence Nationale de la Recherche (ANR) pour le soutien financier apporté au projet CHICA-AI dans le cadre de l'Appel à Projet Générique 2023, ainsi que Renaissance Learning, qui finance le reste du projet et permet l'existence de la plateforme Lalilo.

# Références


CASAZZA M. E. (1993). Using a model of direct instruction to teach summary writing in a college reading class. *Journal of Reading*, **37**(3), 202–208.

CÈBE S., GOIGOUX R. & THOMAZET S. (2004). Enseigner la compréhension. Principes didactiques, exemples de tâches et d'activités. In *Lire écrire, un plaisir retrouvé*. MEN-DESCO. HAL : hal-00922482.

COLMANT M. & LE CAM M. (2017). PIRLS 2016. DOI : 10.48464/ni-17-24, HAL : halshs-03846903.

KOCH G. (2006). *Intraclass Correlation Coefficient*. DOI : 10.1002/0471667196.ess1275.pub2.

OCDE (2016). *Résultats du PISA 2015 (Volume I) : L'excellence et l'équité dans l'éducation*. PISA, Éditions OCDE. DOI : https ://doi.org/10.1787/9789264267534-fr.

OCDE (2019). *Résultats du PISA 2018 (Volume I) : Savoirs et savoir-faire des élèves*. PISA, Éditions OCDE. DOI : https ://doi.org/10.1787/ec30bc50-fr.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Apprentissage par renforcement contraint guidé par un graphe de connaissances pour personnaliser les parcours d'apprentissage

Rania Ait Chabane[1,2] Armelle Brun[1] Azim Roussanaly[1]

(1) Université de Lorraine, LORIA (2) Stellia.ai

rania.ait-chabane@loria.fr, armelle.brun@loria.fr

RÉSUMÉ

Ce travail présente une architecture d'apprentissage adaptatif combinant graphes de connaissances enrichis et contraintes pédagogiques dans un cadre d'apprentissage par renforcement. Le graphe est construit à partir de ressources expertes (ex. manuel scolaire) et enrichi automatiquement par un modèle de langage pour compléter les relations et inférer des contraintes. Un module de knowledge tracing estime la progression de l'apprenant vers un objectif pédagogique donné. Un agent de renforcement, entraîné en environnement simulé, recommande des activités en maximisant la progression attendue tout en respectant les contraintes. Cette approche vise à renforcer la pertinence, la diversité et l'explicabilité des parcours proposés. Une évaluation sur des jeux de données réels est prévue en travaux futurs.

ABSTRACT

**Constrained Reinforcement Learning Guided by a Knowledge Graph for Personalized Learning Pathways**

This work presents an adaptive learning architecture that integrates enriched knowledge graphs and pedagogical constraints into a reinforcement learning framework. The graph is built from expert resources (e.g., textbooks) and enriched using a language model to infer semantic relations and constraints. A knowledge tracing module estimates the learner's progress toward a given objective. A reinforcement learning agent, trained in a simulated environment, recommends optimal activities by maximizing expected learning gains while respecting constraints. This approach aims to improve the relevance, diversity, and explainability of personalized learning paths. Future work will focus on evaluation using real-world educational datasets.

MOTS-CLÉS : Apprentissage adaptatif, Graphes de connaissances, Apprentissage par renforcement.

KEYWORDS: Adaptive learning, Knowledge graphs, Reinforcement learning.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

## 1 Introduction

L'émergence des plateformes d'apprentissage en ligne a profondément modifié les pratiques éducatives, en ouvrant la voie à une personnalisation des parcours d'apprentissage. Le Knowledge Tracing (KT) est un levier central de cette personnalisation (Abdelrahman *et al.*, 2023). Il modélise l'évolution du niveau de maîtrise des concepts à partir des traces d'apprentissage. Les premières approches bayésiennes (Corbett & Anderson, 1994; Käser *et al.*, 2017) adoptent une approche probabiliste. Le

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 17–20.

Deep KT (Piech *et al.*, 2015) a ensuite introduit des réseaux neuronaux récurrents pour capturer des dépendances complexes au sein des séquences d'interactions. Plus récemment, le Self-Attentive KT (Pandey & Karypis, 2019) utilise des mécanismes d'attention pour pondérer l'importance des interactions passées. Toutefois, ces modèles restent limités par leur faible interprétabilité et par une prise en compte partielle du contexte. Le Graph-based KT (GKT) proposé par (Nakagawa *et al.*, 2019) adresse ce point en apprenant une structure sémantique implicite, c'est-à-dire à partir des interactions, tout en prédisant le niveau de maîtrise grâce aux *Graph Neural Network* (GNN). En parallèle, la modélisation explicite du domaine d'apprentissage via des graphes de connaissance (*Knowledge Graphs*-KG) s'est imposée comme une solution pour représenter les concepts, les compétences, les activités, leurs relations et leurs dépendances (Ji *et al.*, 2022). Ces graphes, en apportant une connaissance experte structurée, guident les modèles prédictifs et renforcent à la fois leurs performances et leur explicabilité (Xian *et al.*, 2019). Des travaux récents mobilisent aussi les *Large Language Models* (LLMs) pour construire ou enrichir des KG éducatifs. Ainsi, (Jhajj *et al.*, 2024) utilisent GPT-4 pour générer des EduKG alignés sur des objectifs pédagogiques, et (Abu-Rasheed *et al.*, 2025) utilisent un LLM pour la complétion de graphes de parcours universitaires. Enfin, l'élément clé de la personnalisation est l'algorithme dit d'*adaptive learning* chargé d'identifier l'activité à proposer à l'apprenant à chaque étape de son parcours. Plusieurs approches sont explorées : bandits multi-bras (Mui *et al.*, 2021), classification non supervisée (Chen & Huang, 2023), ou systèmes à base de règles (Kolekar *et al.*, 2019). Toutefois, le *Reinforcement Learning* (RL) est privilégié grâce à sa capacité à optimiser séquentiellement les décisions sur le long terme selon l'évolution de l'état de l'apprenant. Des architectures hybrides combinant KG, KT et RL ont récemment vu le jour. Par exemple, TGKT-RL (Chen *et al.*, 2023) s'appuie sur un graphe question-compétence et un module KT fondé sur GAT+TCN pour guider l'apprenant vers une maîtrise ciblée.

Ces approches présentent deux limites majeures : une sémantique des graphes de connaissances pauvre, et l'absence de contraintes pédagogiques encadrant l'apprentissage par renforcement. Pour y répondre, nous proposons (i) d'exploiter un KG, issu d'une expertise humaine et enrichi sémantiquement par LLM et (ii) d'intégrer des contraintes pédagogiques pour guider l'exploration dans le processus RL.

# 2 Contribution

L'architecture proposée dans ce projet de recherche repose sur trois étapes, présentées Figure 1 :

**L'étape 1** vise à une représentation sémantique riche du domaine pour adapter l'apprentissage de manière pertinente. Elle repose sur la construction d'un KG à partir d'une source experte, comme un manuel scolaire. Ce choix limite l'intervention d'experts humains, souvent coûteuse et chronophage (Cukurova *et al.*, 2023). Les éléments pédagogiques clés ainsi que leurs relations de précédence sont extraits via des techniques de TAL en suivant la structure hiérarchique du manuel. Les traces d'apprentissage issues d'une plateforme e-learning structurée selon le manuel, sont intégrées au KG. De façon optionnelle, pour pallier d'éventuelles insuffisances dans les connaissances extraites du manuel, un LLM est utilisé pour ajouter des relations sémantiques entre concepts. En complément, le LLM est mobilisé pour inférer automatiquement des contraintes pédagogiques à partir du KG et de l'objectif visé par l'apprenant. Ces contraintes peuvent inclure, par exemple, une modulation des prérequis en fonction du niveau actuel de l'apprenant. Elles sont ensuite intégrées au graphe pour guider plus finement la personnalisation du parcours d'apprentissage.

**L'étape 2** est un module de KT, adapté pour aller au-delà de l'évaluation du niveau de connaissance d'une activité, il permet également d'évaluer le niveau de connaissance d'un objectif pédagogique (par

exemple les dérivées). Le contexte ici est plus riche que dans (Nakagawa *et al.*, 2019) : l'identifiant de l'apprenant, le KG enrichi, et l'objectif pédagogique. Le défi ici sera d'adapter les méthodes de KT sur des objectifs qui ne sont pas des activités.

**L'étape 3** repose sur un agent d'apprentissage par renforcement (RL) entraîné de façon classique en deux temps. Dans un premier temps, l'agent interagit avec un environnement simulé, représenté par un module de Graph-based Knowledge Tracing (GKT) (Nakagawa *et al.*, 2019). Ce dernier utilise la structure du graphe de connaissances enrichi pour estimer, à chaque étape, l'évolution du niveau de maîtrise de l'apprenant en réponse à une activité donnée. Le GKT joue donc ici le rôle d'environnement simulé, dans lequel les conséquences pédagogiques des actions peuvent être prédites avant d'être proposées à l'apprenant réel.

La politique de l'agent d'apprentissage par renforcement (RL) vise à sélectionner l'activité pédagogique la plus pertinente en s'appuyant sur un réseau de neurones à graphes (GNN), exploité à partir de la structure sémantique du graphe de connaissances (KG) et des contraintes pédagogiques inférées. Il est prévu de formaliser les contraintes (étape 1) de manière explicite. D'une part, les contraintes fortes, telles que le respect des prérequis, doivent être directement intégrées à la politique de l'agent en restreignant l'espace des actions admissibles. D'autre part, les contraintes faibles, comme la promotion de la diversité des types d'activités, interviennent dans la fonction de récompense, via une modulation de sa valeur. Cette récompense est définie comme la variation estimée du niveau de maîtrise de l'objectif d'apprentissage, évaluée suite à l'exécution simulée d'une activité par l'agent.

Dans un second temps, l'agent communique l'activité choisie à l'apprenant réel. Une fois l'activité réalisée, la nouvelle trace d'apprentissage est intégrée au KG. Ce dernier évolue dynamiquement en intégrant les nouvelles connaissances observées chez l'apprenant, ce qui lui permet de jouer un double rôle : à la fois représentation du domaine et le suivi temporel de la progression individuelle de l'apprenant. Un nouveau cycle d'interaction est alors relancé jusqu'à atteinte de l'objectif pédagogique.

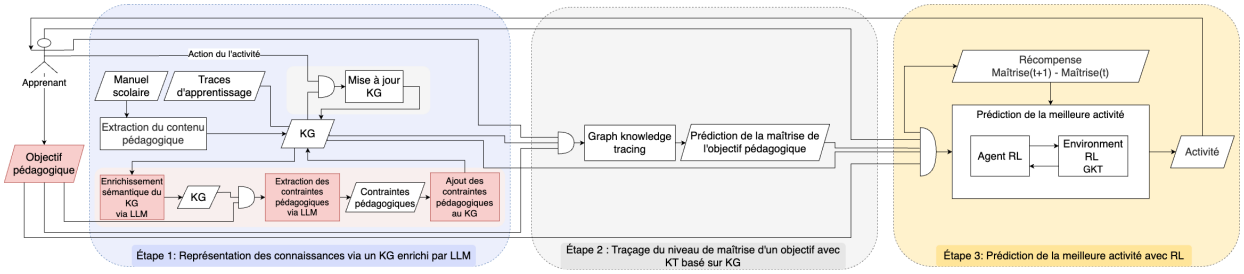



FIGURE 1 – Architecture générale du système proposé.

# 3 Conclusion

Les prochaines étapes concerneront l'implémentation et l'évaluation de l'architecture à partir de jeux de données comme ASSISTments. Les performances seront évaluées selon trois critères : (1) l'amélioration du niveau de maîtrise, mesurée par l'AUC et le RMSE entre les prédictions du module de Knowledge Tracing et les réponses réelles des apprenants ; (2) la diversité des parcours recommandés, évaluée par l'entropie moyenne des types d'activités proposées et la distance de Jaccard entre séquences d'apprenants ; (3) l'explicabilité des recommandations, analysée qualitativement par des experts et quantitativement via la longueur moyenne des chaînes de raisonnement extraites du graphe.

# Références


ABDELRAHMAN G., WANG Q. & NUNES B. (2023). Knowledge tracing : A survey. *ACM Computing Surveys*, **55**(11), 1–37.

ABU-RASHEED H. *et al.* (2025). Llm-assisted knowledge graph completion for curriculum and domain modelling in personalized higher education recommendations. *arXiv preprint arXiv :2501.12300*.

CHEN L. & HUANG H. (2023). Adaptive e-learning system based on learner portraits and knowledge graph. In *2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA)*, volume 3.

CHEN Z., WANG Q. & NUNES B. (2023). Tgkt-based personalized learning path recommendation with reinforcement learning. In *International Conference on Knowledge Science, Engineering and Management* : Springer Nature Switzerland.

CORBETT A. T. & ANDERSON J. R. (1994). Knowledge tracing : Modeling the acquisition of procedural knowledge. In *User Modeling and User-Adapted Interaction*, volume 4, p. 253–278.

CUKUROVA M. *et al.* (2023). Adoption of artificial intelligence in schools : Unveiling factors influencing teachers' engagement. In *International conference on artificial intelligence in education* : Springer Nature Switzerland.

JHAJJ G. *et al.* (2024). Educational knowledge graph creation and augmentation via llms. In *International Conference on Intelligent Tutoring Systems* : Springer Nature Switzerland.

JI S., PAN S., CAMBRIA E., MARTTINEN P. & YU P. S. (2022). A survey on knowledge graphs : Representation, acquisition and applications. *IEEE Transactions on Neural Networks and Learning Systems*, **33**(2), 494–514.

KÄSER T., KLINGLER S., SCHWING A. & GROSS M. (2017). Dynamic bayesian networks for student modeling. *IEEE Transactions on Learning Technologies*, **10**(4), 450–462.

KOLEKAR S. V., PAI R. M. & PAI M. M. (2019). Rule based adaptive user interface for adaptive e-learning system. *Education and Information Technologies*, **24**, 613–641.

MUI J., LIN F. & DEWAN M. A. A. (2021). Multi-armed bandit algorithms for adaptive learning : A survey. In *International Conference on Artificial Intelligence in Education* : Springer International Publishing.

NAKAGAWA H., IWASAWA Y. & MATSUO Y. (2019). Graph-based knowledge tracing : Modeling student proficiency using graph neural network. In *IEEE/WIC/ACM International Conference on Web Intelligence*.

PANDEY S. & KARYPIS G. (2019). A self-attentive model for knowledge tracing. *arXiv preprint arXiv :1907.06837*.

PIECH C., BASSEN J., HUANG J., GANGULI S., SAHAMI M., GUIBAS L. & SOHL-DICKSTEIN J. (2015). Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, volume 28.

XIAN Y., FU Z., WANG S., ZHANG Z., YAO Y. & SUN S. (2019). Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Découverte de l'intelligence artificielle par des directeurs et directrices d'école primaire : une étude de cas dans deux circonscriptions marseillaises

Hervé Allesant[1] Ismail Badache[2] Maria Impedovo[3]
(1) Aix-Marseille Université, Académie d'Aix-Marseille, Marseille, France
(2) Aix-Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France
(3) Aix Marseille Université, ADEF, Marseille, France
herve.allesant@ac-aix-marseille.fr, ismail.badache@univ-amu.fr,
maria.impedovo@univ-amu.fr

## RÉSUMÉ

Cet article présente une étude préliminaire sur l'engagement de directeurs et directrices d'écoles primaires à Marseille (France) vis-à-vis de l'intelligence artificielle (IA), en particulier de l'IA générative, dans le cadre du numérique éducatif et de la transformation digitale de l'école. L'étude analyse un atelier de formation visant à introduire l'évolution historique de l'IA, ses fondements ainsi que ses applications pédagogiques. Les résultats, issus de questionnaires administrés avant et après l'intervention, montrent que, malgré une connaissance initiale limitée des technologies d'IA, les participants ont manifesté un intérêt croissant pour l'exploration de ces outils, principalement dans une optique de gain de temps, tout en conservant une méfiance marquée. L'étude souligne la nécessité de formations contextualisées, combinant connaissances et compétences techno-pédagogiques en IA et réflexion critique, et appelle à une prise en compte des enjeux éthiques et des cadres de gouvernance pour une intégration responsable de l'IA en éducation.

## ABSTRACT

**Primary School Principals Discovering Artificial Intelligence : A Case Study in Two Districts of Marseille.**

This paper presents a preliminary study on the engagement of primary school principals in Marseille (France) with artificial intelligence (AI), particularly generative AI, in the context of educational digitalization and the digital transformation of schools. The study analyzes a training workshop designed to introduce the historical evolution of AI, its foundations, and its pedagogical applications. The results, drawn from questionnaires administered before and after the intervention, reveal that despite participants' initially limited knowledge of AI technologies, they demonstrated growing interest in exploring these tools, primarily motivated by time-saving goals, while maintaining marked skepticism. The study underscores the need for contextualized training that combines knowledge and techno-pedagogical skills in AI with critical reflection. It also calls for addressing ethical considerations and governance frameworks to ensure responsible integration of AI in education.

MOTS-CLÉS : IA à l'école, numérique éducatif, formation à l'IA, transformation numérique.

KEYWORDS: AI in schools, educational digital, AI training, digital transformation..

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 21–28.

# 1 Introduction

L'intérêt du grand public autour des intelligences artificielles (IA) génératives (IAg) a suscité des questionnements de la part de plusieurs directeurs et directrices d'école travaillant sur le territoire d'intervention d'un des auteurs de cet article (en lien avec l'académie d'Aix-Marseille et la région PACA Provence-Alpes-Côte d'Azur en France). Nous avons pu constater que les utilisations de l'IAg pour un accompagnement des gestes des enseignants sont souvent orientées vers le second degré, présentées par des spécialistes, pour des spécialistes, issus du côté technique : informaticiens, ingénieurs, professeurs du secondaire ou d'université. Dans notre cas, l'enseignant du primaire, de par sa polyvalence, souhaite obtenir lors d'une formation, des informations utilisables et acceptables (Tricot *et al.*, 2003; Davis, 1989) pour accompagner ses gestes professionnels rapidement.

Partant de cette constatation, et pour répondre aux besoins exprimés par les agents, un atelier de formation a donc été proposé aux directeurs et directrices d'école des circonscriptions du premier degré de Huveaune et St Barnabé, à Marseille (France). Les ateliers étaient proposés par l'Enseignant Ressource en Usages du Numérique (ERUN) [1].

Cet article a pour objectif de recueillir et d'analyser l'impact d'un atelier de formation à la découverte de l'IA sur les représentations et les pratiques professionnelles des directeurs d'écoles publiques du premier degré.. Il s'agit notamment d'étudier l'évolution de leurs idées reçues concernant les IA génératives, ainsi que les usages qu'ils peuvent en faire dans l'exercice quotidien de leurs fonctions de pilotage pédagogique, avant et après la formation. Ce recueil de données constitue une étape préliminaire d'une recherche plus large, visant à identifier les pratiques pédagogiques et managériales efficaces, et à orienter la conception de formations futures sur l'IA. L'enjeu est d'apporter des réponses concrètes et adaptées aux besoins des équipes de direction et des enseignants du premier degré, dans le cadre du système éducatif français ou d'autres systèmes éducatifs équivalents.

L'organisation de cet article se structure comme suit : dans un premier temps, nous présentons quelques dispositifs de formation mis en place, ainsi que notre démarche vers des ateliers contextualisés sur l'IA en éducation destiné spécifiquement aux directeurs et directrices d'écoles à Marseille. Dans un second temps, nous détaillons la méthodologie de l'enquête, les modalités de collecte des données et nous analysons en discutant quelques résultats préliminaires de l'étude. Enfin, nous concluons cette recherche exploratoire préliminaire et présentons les perspectives.

# 2 Dispositifs de formation en IA : vers des ateliers contextualisés

Axel Jean, chef du bureau du soutien à l'innovation numérique et à la recherche appliquée à la Direction du numérique pour l'Éducation (DNE) au ministère de l'Éducation Nationale (ÉN), de l'Enseignement supérieur et de la Recherche, évoque l'urgence de mettre en place une acculturation à l'IA au sein de l'ÉN. Il s'exprime à ce sujet lors d'un entretien intitulé « *Peut-on mettre l'IA au service de l'éducation ?* » [2], en compagnie d'Orianne Ledroit, déléguée générale d'EdTech France et enseignante à Sciences Po Paris, ainsi que de Mickaël Bertrand, enseignant d'histoire-géographie et d'EMC en lycée et en milieu carcéral (académie de Dijon). Même si des enseignants, dans une dizaine de pays (Miao & Shiohira, 2022) ont commencé à construire des programmes d'acculturation à l'IA (Chiu *et al.*, 2024), ces programmes tiennent finalement peu compte de la spécificité des élèves du primaire (Yim, 2023) alors que l'enjeu de la connaissance de l'IA pour des élèves de cet âge est capitale (Chiu *et al.*, 2024; Almatrafi *et al.*, 2024) pour leur carrière professionnelle (Li, 2024), tout en réduisant la fracture numérique, quel que soit leur milieu d'origine. (Luckin & Holmes, 2016).

1. https://primabord.eduscol.education.fr/les-erun-acteurs-du-numerique-educatif
2. https://www.radiofrance.fr/franceculture/podcasts/etre-et-savoir/education-et-ia-ou-allons-nous-7616037

En France, l'une des solutions proposées par l'Éducation Nationale, orientée vers ses agents, est le Café IA[3]. Le concept de Café IA s'inscrit dans la continuité des travaux du Conseil national du numérique[4], notamment suite à la publication en février 2021 du rapport *Civilisation numérique. Ouvrons le débat !*[5] et du rapport *IA, notre ambition pour la France*[6] en mars 2024. Ce dernier recommandait la création d'espaces d'échanges pour permettre une appropriation collective de l'IA. Ainsi, le 21 mai 2024, le président de la République a chargé le Conseil de structurer le projet Café IA. À partir de l'idée initiale du Café IA, et en nous appuyant sur les enjeux de formation identifiés par le GTnum GTnum[7] (Allouche, 2023, 2025), nous avons conçu un atelier contextualisé autour de l'IA, adoptant une approche techno-pédagogique spécifiquement pensée pour les directeurs et directrices d'écoles primaires. Cet atelier s'articule autour de trois axes principaux :

- **Historique** : Une mise en perspective historique de l'IA, s'appuyant sur les ressources proposées par *That's AI*[8], retrace l'évolution des IA depuis les premières victoires en jeu de dames jusqu'aux performances spectaculaires dans le jeu de go. Cette progression illustre la transition entre deux paradigmes majeurs : d'une part, l'IA symbolique, fondée sur des algorithmes explicites sur la base des règles, et d'autre part, l'IA connexionniste, reposant sur l'apprentissage automatique à partir de données massives.
- **Technique** : Une présentation de quelques outils permettant l'expérimentation avec des IA, tout en soulignant les limites actuelles liées à leur usage en contexte scolaire. En particulier, les grands modèles de langage (LLM, en anglais *Large Language Models*) ne sont, à ce jour, pas pleinement exploitables en classe en raison des contraintes réglementaires imposées par le Règlement Général sur la Protection des Données (RGPD), telles qu'interprétées par la CNIL en 2025[9].
- **Pratique** : Un temps dédié à la mise en pratique, visant à expliciter les principes de l'ingénierie de prompt dans l'interaction avec les modèles de langage. À cette occasion, plusieurs acronymes méthodologiques sont proposés aux enseignants, tels que le modèle ACTIF[10] — **A**ction (la tâche à effectuer), **C**ontexte (exemples, modèles et contraintes), **T**onalité (style attendu), **I**dentité (rôle attribué à l'IA), **F**ormat (forme du résultat attendu) — afin de formaliser et structurer les requêtes (prompts) adressées à un LLM de manière claire, efficace et pédagogique.

# 3 Méthodologie de l'étude : Enquête-Diagnostic

Cette étude s'inscrit dans une démarche de recherche descriptive et exploratoire, reposant sur une méthode mixte articulée en deux phases complémentaires : 1) Une phase quantitative menée en amont et à l'issue de l'atelier de formation (pré-test et post-test), visant à réaliser une enquête-diagnostic préliminaire portant sur les représentations, les pratiques et les besoins en matière d'IA dans le contexte professionnel des directions d'école du premier degré à Marseille, cette démarche a également, en post-test, pour objectif d'analyser les évolutions potentielles induites par l'atelier de sensibilisation, ainsi que d'identifier les besoins spécifiques en vue de la conception d'éventuelles actions de formation complémentaires ; 2) Une phase qualitative conduite à l'issue de l'atelier visant à avoir des réponses claires autour de l'évoltion des usages de l'IA.

3. https://cafeia.org/grands-principes-cafe-ia/
4. https://cnnumerique.fr/
5. Rapport cnnumerique - Civilisation numérique. Ouvrons le débat !
6. 25 recommandations pour l'IA en France.
7. Groupes Thématiques numériques
8. That's AI - L'histoire de l'IA
9. IA et RGPD : la CNIL publie ses nouvelles recommandations pour accompagner une innovation responsable
10. https://actif.numedu.org/

Le recueil des données a été mené dans le cadre d'un atelier de formation organisé pendant les temps de décharge de direction, ces derniers constituant un moment privilégié pour la participation volontaire des directeurs et directrices, compte tenu de leur double mission de gestion et d'enseignement. L'atelier a rassemblé 40 participants (directeurs et directrices d'écoles à Marseille) et a duré 2 heures. Parmi eux, 22 ont répondu au questionnaire pré-atelier, et 17 ont répondu au questionnaire post-atelier.

Cette approche relève de la méthodologie de l'enquête-diagnostic telle que définie dans les recherches en sciences de l'éducation (Van der Maren, 1996; Savoie-Zajc, 2018), c'est-à-dire une investigation préliminaire qui permet de dresser un état des lieux avant une recherche plus approfondie afin de mettre en œuvre des actions ou des dispositifs éducatifs.

## 3.1 Collecte des données

Afin de documenter l'évolution des représentations, ressentis et usages liés à l'IA dans le contexte professionnel des directions d'école, deux questionnaires en ligne, anonymes, ont été administrés aux participants : l'un en amont de l'atelier, l'autre dans les jours suivant sa tenue.

Conformément aux principes de la triangulation méthodologique (Denzin, 2017), le premier questionnaire, à dominante quantitative, comprenait des items à choix fermés et multiples. Il visait à : a) mesurer les compétences numériques autodéclarées ; b) évaluer le niveau de familiarité avec les IA génératives ; c) apprécier l'évolution des perceptions à la suite de l'atelier ; d) analyser l'impact perçu des trois volets abordés lors de la formation (historique, technique, pratique). Le second questionnaire, de nature qualitative, avait pour objectif de recueillir des données plus approfondies sur les usages effectifs de l'IA par les directeurs et directrices, ainsi que sur les tâches professionnelles qu'ils envisageraient de déléguer ou non à ces technologies.

Cette approche méthodologique mixte s'inscrit dans une perspective compréhensive (Miles & Huberman, 2003), en articulant données objectives et subjectives pour appréhender la diversité des postures vis-à-vis de l'IA dans le premier degré.

TABLE 1 – Analyse comparative avant/après l'atelier IA

| Catégorie | Avant l'atelier | Après l'atelier | Réflexivité |
|---|---|---|---|
| Utilisation de l'IA | **50%** n'utilisaient pas l'IA. Usages limités, principalement pour la rédaction et la traduction. | Diversification des usages (rédaction, création visuelle, synthèse). **Gain de temps** perçu (+30 min en moyenne) | L'atelier a encouragé l'expérimentation et aller vers d'autres usages. |
| Compétences perçues | Note moyenne : **2.8/4** Connaissance faible en IA | Confiance accrue suite à l'**amélioration des connaissances**. | Renforcement des compétences techniques. |
| Réflexions critiques | Peu d'éthique | Prise de **conscience** et volonté de **former les élèves**. | Encouragement à la formation pour un usage responsable |
| Réserves et craintes | Confidentialité (**35%**) et Biais (**25%**) | Craintes persistantes avec vérification systématique | Limites éthiques restent prégnantes car **manque d'outils IA** véritablement **adaptés**. |
| Tâches déléguées | Peu de tâches identifiées | Nouvelles délégations : gestion des mails, planning, rédaction de documents, ect. | Identification de cas d'usage concrets et professionnels |
| Tâches non délégables | Interactions humaines, données sensibles. | Idem, avec une insistance sur la **supervision humaine**. | Conscience ou juste par crainte des limites de la délégation. |
| Impact perçu | **70%** utilisation future | **100%** adoption prévue avec une efficacité accrue | Transformation en outil pratique au service de la pédagogie |

## 3.2 Analyse et discussion des résultats avant/après l'atelier

Le tableau 1 présente une analyse comparative avant et après l'atelier IA, mettant en lumière les évolutions observées dans différentes catégories liées à l'usage de l'IA. Il se structure autour de 7 axes principaux (colonne Catégorie). Chaque catégorie est évaluée selon 3 volets : avant l'atelier, après l'atelier, et les réflexivités ou conclusions tirées de cette expérience. Cette analyse vise à identifier

les changements dans les comportements, les perceptions et les pratiques des participants, tout en soulignant les bénéfices et les défis associés à l'intégration de l'IA dans leurs activités professionnelles.

**Avant l'atelier.** Bien que chatGPT ait popularisé les IA génératives depuis novembre 2022, à la date de l'atelier, en janvier 2025, plus de la moitié des participants n'avait jamais utilisé une IA générative (voir Tableau 1). Les directeurs du premier degré se sont évalués comme peu compétents face à l'outil informatique (39%) et seulement 4% se classent dans la catégorie la plus haute, alors qu'une majorité des outils de pilotage et de direction sont désormais numériques (Glomeron, 2015). L'utilisation des TICE reste un domaine peu enseigné dans la formation initiale des enseignants, seulement 34% ayant profité d'un enseignement dans ce domaine, et 25% déclarant avoir un sentiment de préparation autour du numérique éducatif (DEPP, 2025). Dans les participants qui avaient manipulé une IA au préalable, une moitié l'avait utilisée pour rechercher des informations, un usage peu pertinent car on sait le travers que peuvent avoir les LLM quant à la production d'hallucinations (Sun *et al.*, 2024).

TABLE 2 – Quelques résultats avant l'atelier

| Indicateur | Observation |
|---|---|
| Non je n'avais pas utilisé l'IA avant l'atelier | 11 participants |
| Recherche et collecte de contenus | 11 participants |
| Créations de supports pédagogiques pour les cours (génération de contenus, résumés etc.) | 4 participants |
| Elaboration de devoirs à donner aux élèves | 1 participant |
| Traductions automatiques | 2 participants |
| Autres usages pédagogiques (ex. Coding, etc. | 1 participant |
| Génération de textes (rapports, programmations etc.) | 3 participants |
| Création d'images | 3 participants |
| Exploration des potentialités de l'IA | 2 participants |

**Après l'atelier.** À la suite de l'atelier, une grande majorité des participants a commencé à manipuler l'IA, et ce, pour des usages plus variés qu'avant l'atelier (voir le tableau 3). Les réponses ne sont pas univoques : plus de réponses sont possibles pour chaque demande. Certains usages pourraient sembler incongrus dans un cadre pédagogique, car nous avions notamment abordé la création de musiques avec le site suno.ai pour générer une chanson reprenant le contenu de certaines leçons. A partir d'un prompt, cette IA générative peut produire un texte et le mettre en musique, ou bien l'on peut donner un texte, qui sera mis en musique selon le style que l'on aura choisi. Pendant l'atelier, le poème de Maurice Carême "Le chat et le soleil" [11] a été mis en musique dans un style de rap urbain.

TABLE 3 – Quelques résultats après l'atelier

| Indicateur | Observation |
|---|---|
| Non je n'ai pas utilisé l'IA suite à l'atelier | 2 participants |
| Recherche et collecte de contenus (général) | 9 participants |
| Création de supports pédagogiques pour les cours (génération de contenus, résumés, etc.) | 10 participants |
| Elaboration de devoirs à donner aux élèves | 2 participants |
| Conception de tests ou d'évaluations | 3 participants |
| Traductions automatiques | 4 participants |
| Autres usages pédagogiques (ex. Coding etc.) | 3 participants |
| Génération de textes (rapports, programmations, etc.) | 10 participants |
| Création d'images | 7 participants |
| Génération de vidéos | 1 participant |
| Génération de sons/musiques | 6 participants |
| Explorations des potentialités de l'IA | 7 participants |
| Support pédagogique pour la résolution de problèmes | 1 participant |

## 3.3 Réponses autour des usages

La plupart des réponses autour des usages de l'IA après l'atelier évoquent le gain de temps que l'outil leur apporte, par exemple pour la rédaction de certains mails délicats. Une réponse en particulier évoque le cas récent de tensions autour de l'éducation à la vie affective et relationnelle (EVAR)

11. https://genius.com/Maurice-careme-le-chat-et-le-soleil-annotated

l'éducation à la sexualité n'étant pas abordée dans le premier degré, le sigle étant EVARS à partir du collège, cela avait créé des incompréhensions et des inquiétudes. Des collectifs de parents avait mené une campagne pour manifester leur mécontentement autour de cet enseignement, et les directeurs ont été invités à communiquer auprès des parents pour les rassurer. Une directrice témoigne :

***Extrait 1 :*** *"Rédaction d'un courrier aux parents d'élèves suite à leurs nombreuses inquiétudes par rapport aux nouveaux programmes d'éducation à la vie affective et relationnelle. Je me suis inspirée de certaines tournures de phrases (pas utilisées en totalité). Je ne sais pas le temps gagné."*

Plusieurs réponses évoquent la rapidité de l'écriture des documents de synthèses de réunions, d'équipes éducatives, tout en évoquant la crainte de se voir remplacer.

***Extrait 2 :*** *"Je suis tentée de demander la rédaction de rapports, mais je préfère mettre en forme mes propres notes, j'ai peur de constater que la machine peut me remplacer sans problème !"*

Plusieurs directeurs veulent garder certaines tâches, répétitives ou chronophages et ce, même si on leur garantit que l'IA pourrait faire aussi bien qu'eux. À la question "*Quelles tâches ne confieriez-vous pas à l'IA, même avec la garantie d'un travail bien fait ?*"

***Extrait 3 :*** *"Tout ce qui touche à la communication directe, émotions, dimension humaine de la tâche. Ce doit être un outil facilitateur, une aide pour gagner du temps, mais toujours supervisée par le dirlo"*
***Extrait 4 :*** *"Commentaires concernant le travail des élèves (cahier de réussite)"*

La plupart des participants évoquent le fait que l'IA n'est qu'un assistant, et ne doit pas être considérée comme autre chose qu'un conseiller ou une aide. Il y a un rapport de fascination et de méfiance face à cet outil, qui implique que le directeur souhaite quand même superviser les productions proposées :

***Extrait 5 :*** *"Ce doit être un outil facilitateur, une aide pour gagner du temps, mais toujours supervisée par le dirlo"*
***Extrait 6 :*** *"Pour moi, l'IA ne fait pas à ma place. Elle m'aide."*
***Extrait 7 :*** *"L'utilisation de l'IA pour le moment est une AIDE, un "autre regard" sur une situation, un document, un courrier à rédiger : a. demande de listing de questions pouvant être posées lors d'un entretien de poste PEP MEG ; b. demande d'informations concernant le dispositif Marseille en Grand sur Marseille (pas seulement les écoles) ; c. résumé de textes, documents, compte-rendus de conseil d'école....etc (mon point faible est de résumer en quelques lignes, ou mots) ; d. résumé et analyse des résultats des évaluations nationales CP de notre école.*

# 4 Conclusion

L'analyse des résultats met en évidence un décalage important entre les pratiques professionnelles et la préparation réelle des directeurs du premier degré, peu formés et peu confiants en leurs compétences numériques. La découverte de l'IA générative à l'occasion de l'atelier élargit leurs usages, orientés vers un gain de temps dans la rédaction de documents. Toutefois, une méfiance demeure : l'IA est perçue comme un outil d'assistance, qui doit être supervisé. Cette posture ambivalente, entre ouverture pragmatique et vigilance éthique, souligne l'importance d'accompagner l'intégration des IA dans les pratiques éducatives par des formations ciblées, entre compétences techniques et réflexion critique sur les enjeux liés à leur usage. L'adoption massive de l'IA par les élèves, simultanément à sa découverte par les enseignants demande du temps de manipulation et des formations. Plus nous attendrons pour accompagner un usage raisonné et constructif , plus l'écart sera difficile à combler entre les habitudes prises par les élèves et l'accompagnement aux usages que nous devrons mettre en place. Une première limite pourra être relevée sur les résultats : les ateliers se tenaient sur temps de décharges. Aucun directeur de petites structures, qui pourraient être les plus intéressés par une assistance basée sur l'IA. Seule une moitié des 80 directeurs a participé aux ateliers et tous n'ont

pas répondu aux questionnaires. Il sera donc pertinent de faire un inventaire à plus grande échelle, visant l'exhaustivité. Une autre limite peut être évoquée : les absents peuvent se retrouver dans deux catégories : ceux pensant maitriser le sujet, ou au contraire, ceux ne voyant pas l'IA comme un sujet relevant de leur développement professionnel.

Cet article, malgré certaines limites, constitue une étude nécessaire sur les usages et perceptions de l'IAg dans le premier degré. Il apporte un éclairage concret sur un champ encore peu exploré, soulignant l'importance d'une compréhension fine des réalités de terrain pour poser les bonnes questions et engager la communauté éducative dans une réflexion responsable sur l'intégration de l'IA en éducation.

# Références


ALLOUCHE E. (2023). IA génératives et ingénierie pédagogique : le prompting, pistes de travail et applications. https://edunumrech.hypotheses.org/9934. Publié dans *Éducation, numérique et recherche*.

ALLOUCHE E. (2025). Intelligence artificielle et éducation : apports de la recherche et enjeux pour les politiques publiques. https://edunumrech.hypotheses.org/13849. Publié dans *Éducation, numérique et recherche*.

ALMATRAFI O., JOHRI A. & LEE H. (2024). A systematic review of ai literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). *Computers and Education Open*, **6**, 100173. DOI : 10.1016/j.caeo.2024.100173.

BADACHE I. & BELLET P. (2024). Intelligence artificielle : usage pédagogique et esprit critique. In *16ème édition du colloque Interactions Multimodales Par ÉCran, IMPEC 2024*. https://hal.science/hal-04659335v1.

CHIU T. K., AHMAD Z., ISMAILOV M. & SANUSI I. T. (2024). What are artificial intelligence literacy and competency ? a comprehensive framework to support them. *Computers and Education Open*, **6**, 100171. DOI : 10.1016/j.caeo.2024.100171.

DAVIS F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, p. 319–340. DOI : 10.2307/249008.

DENZIN N. K. (2017). *The research act : A theoretical introduction to sociological methods*. Routledge. DOI : 10.4324/9781315134543.

GLOMERON F. (2015). L'action du directeur d'école : entre interface et coopération au sein de différents réseaux fonctionnels. In *Biennale internationale de l'éducation et de la formation et des pratiques professionnelles*. https://hal.science/hal-01179860/.

LI L. (2024). Reskilling and upskilling the future-ready workforce for industry 4.0 and beyond. *Information Systems Frontiers*, **26**(5), 1697–1712. DOI : 10.1007/s10796-022-10308-y.

LUCKIN R. & HOLMES W. (2016). Intelligence unleashed : An argument for ai in education. *UCL Knowledge Lab*. https://discovery.ucl.ac.uk/id/eprint/1475756/.

MIAO F. & SHIOHIRA K. (2022). K-12 ai curricula. A mapping of government-endorsed ai curricula. *UNESCO Publishing*, **3**, 60. DOI : 10.54675/ELYF6010.

MILES M. B. & HUBERMAN A. M. (2003). *Analyse des données qualitatives*. De Boeck Supérieur. https://books.google.fr/books?id=AQHRyJ1AiPEC.

SAVOIE-ZAJC L. (2018). La recherche qualitative/interprétative. *La recherche en éducation : étapes et approches*, **4**, 191–217. DOI : 10.1515/9782760639331-009.

SUN Y., SHENG D., ZHOU Z. & WU Y. (2024). AI hallucination : towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, **11**(1), 1–14. DOI : 10.1057/s41599-024-03811-x.

TRICOT A., PLÉGAT-SOUTJIS F., CAMPS J.-F., AMIEL A., LUTZ G. & MORCILLO A. (2003). Utilité, utilisabilité, acceptabilité : interpréter les relations entre trois dimensions de l'évaluation des EIAH. In *Environnements Informatiques pour l'Apprentissage Humain 2003*, p. 391–402 : ATIEF. https://edutice.hal.science/edutice-00000154v1.

VAN DER MAREN J.-M. (1996). *Méthodes de recherche pour l'éducation*. Presses de l'Université de Montréal et de Boeck. DOI : 10.7202/031875ar.

YIM I. H. Y. (2023). Design of artificial intelligence (ai) education for primary schools : arts-based approach. *Istes Books*, p. 65–90. https://book.istes.org/index.php/ib/article/view/5.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Exploration du RAG pour la génération de réponses à des questions en contexte éducatif: étude sur les données SCIQ

Sarah Nouali Ismail Badache Patrice Bellot
Aix-Marseille University, Université de Toulon, CNRS, LIS, Marseille, France
{sarah.nouali, ismail.badache, patrice.bellot}@lis-lab.fr

## RÉSUMÉ

Les systèmes basés sur le RAG (Retrieval-Augmented Generation) sont des systèmes qui optimisent la puissance des grands modèles de langue (LLM, en anglais, *Large Language Models*) avec une recherche d'information (RI) à partir de sources de connaissances externes, sans avoir besoin de réentraîner le modèle. Ce type d'approche est connu pour améliorer les réponses du LLM, en particulier pour répondre à des questions spécifiques à un domaine, et réduire le phénomène d'hallucination constaté avec ces derniers. Dans cet article, nous explorons l'application d'un tel système dans un contexte pédagogique, en utilisant le jeu de données SCIQ (*SCIence Questions*), un ensemble de questions scientifiques à choix multiples de niveau scolaire, qui nous permet d'évaluer la capacité des modèles à fournir des réponses précises, pédagogiques et vérifiables. Nous évaluons les performances du système par rapport à un modèle génératif standard (*Llama3 8b* et *Mistral 7b*) de réponse aux questions et analysons ses forces et ses limites dans un contexte éducatif. La performance la plus élevée en termes de précision a été enregistrée avec l'approche basée sur le RAG (*rag-llama*), qui a permis d'atteindre une précision globalement supérieure par rapport aux autres approches testées.

## ABSTRACT

**Exploring RAG for educational question answering : A study on the SCIQ dataset**

Retrieval-Augmented Generation (RAG) based systems are systems that optimize the power of large language models (LLMs) with information retrieval from external knowledge sources, without the need to re-train the model. This type of approach is known to improve LLM responses, particularly when answering domain-specific questions, and reduce the hallucination phenomenon seen with the latter. In this article, we explore the application of such a system in a pedagogical context, using the SCIQ dataset, set of grade-level multiple-choice scientific questions, which enables us to assess the models' ability to provide accurate, pedagogical and verifiable answers. We evaluate the system's performance against a standard generative question answering model (LLM) and analyze its strengths and limitations in an educational context. The highest performance in terms of accuracy was recorded with the RAG-based approach (*rag-llama*), which achieved an overall higher accuracy than the other approaches tested.

MOTS-CLÉS : système question-réponse, grands modèles de langue, RAG, éducation, SCIQ.

KEYWORDS: question-answer system, large language models, RAG, education, SCIQ.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 29–41.

# 1 Introduction

De nos jours, l'intelligence artificielle (IA) est de plus en plus présente dans de nombreux domaines, et l'éducation n'échappe pas à cette tendance. L'IA est désormais intégrée dans les environnements éducatifs à travers des applications variées : pour l'aide aux élèves via des systèmes de tutorat intelligent, apprentissage ludique (*game-based learning*) ou encore évaluation formative automatisée, pour apporter un soutien aux enseignants avec des outils de détection de plagiat, de curation intelligente de ressources pédagogiques ou d'évaluation sommative automatisée, ou pour fournir une assistance aux institutions et établissements scolaires grâce à des systèmes de gestion des admissions, de planification des cours et des emplois du temps, ou encore de surveillance à distance des examens (Holmes & Tuomi, 2022). Parmi ces usages, les systèmes de tutorat intelligent (STI) comptent parmi les applications d'IA les plus répandues et les mieux financées dans le domaine de l'éducation. Ils proposent des tutoriels informatisés, étape par étape, adaptés à chaque élève, principalement dans des disciplines structurées comme les mathématiques (Holmes & Tuomi, 2022).

Dans ce contexte, les systèmes de réponse automatique aux questions (QA) apparaissent comme particulièrement prometteurs pour accompagner l'apprentissage des élèves et fournir des explications à la demande. Toutefois, les modèles de langage de grande taille (LLMs), bien que puissants, génèrent leurs réponses à partir de connaissances apprises lors de l'entraînement. Cela peut entraîner des réponses erronées, obsolètes, ou encore des phénomènes d'hallucination — une faiblesse bien connue de ces systèmes. Pour pallier ces limites, les systèmes de type génération augmentée par la recherche d'information *Retrieval-Augmented Generation* (RAG) (Lewis *et al.*, 2020) ont été proposés. Ils combinent les capacités génératives des LLMs avec un module de recherche documentaire, permettant de produire des réponses fondées sur des sources externes fiables et vérifiables, telles que des articles scientifiques ou des manuels scolaires. Dans un contexte éducatif, cette capacité à s'appuyer sur des sources factuelles est essentielle. Elle garantit des réponses plus fiables, alignées avec les objectifs pédagogiques définis par les programmes scolaires. De plus, les systèmes RAG offrent une transparence accrue : les documents utilisés pour formuler la réponse peuvent être consultés, ce qui permet aussi bien aux élèves qu'aux enseignants de comprendre le raisonnement sous-jacent. Enfin, ces systèmes peuvent être adaptés à des corpus spécifiques, permettant un alignement précis avec des référentiels pédagogiques ou des supports institutionnels. Ainsi, les systèmes RAG représentent une alternative plus fiable, interprétable et potentiellement mieux adaptée aux exigences de l'enseignement que les modèles génératifs classiques.

Ce travail a pour objectif d'explorer de manière approfondie le potentiel de cette approche dans un contexte éducatif, en s'appuyant sur un ensemble de questions de recherche structurantes qui guident l'analyse et la réflexion.

— **Q1 :** Comment intégrer le RAG dans un environnement de tutorat intelligent, et quelle est son efficacité pour ce type de tâche ?

— **Q2 :** Comment l'utilisation du RAG peut-elle renforcer la confiance des utilisateurs envers les réponses du système dans un contexte éducatif où la véracité est primordiale ?

— **Q3 :** Est-ce que les systèmes basés sur le RAG permettent réellement de fournir des sources ou des supports justifiant les réponses générées ?

— **Q4 :** Les réponses générées par un système RAG sont-elles suffisamment explicites et claires pour favoriser la compréhension des apprenants ?

# 2 IA et RAG en éducation : une vue d'ensemble

La présente section propose une vue d'ensemble des apports de l'intelligence artificielle (IA) en éducation, en articulant deux axes complémentaires. Le premier examine les usages généraux de l'IA dans le domaine éducatif, en s'appuyant sur les évolutions récentes des technologies basées sur les modèles de langage. Le second s'intéresse plus spécifiquement aux systèmes dits RAG, conçus pour pallier certaines limites des LLMs, en particulier le phénomène d'hallucination, en intégrant des mécanismes de récupération d'informations provenant de sources externes fiables.

## 2.1 l'IA et l'éducation

Nous pouvons classer les solutions d'IA appliquées à l'éducation en trois catégories : celles destinées à aider les élèves dans leur apprentissage, celles conçues pour accompagner les enseignants, et enfin celles dédiées au soutien des établissements et institutions scolaires (Holmes & Tuomi, 2022). Dans la catégories des applications pour élèves, nous retrouvons plusieurs types de solutions : les systèmes de tutorat intelligent (Ward *et al.*, 2013; Dolenc *et al.*, 2015; Lane *et al.*, 2013), les environnements d'apprentissage exploratoires, les simulations assistées par l'IA comme les jeux éducatifs numériques (McLaren *et al.*, 2017; Parong *et al.*, 2017; Mayer, 2019), les agents conversationnels (Paschoal *et al.*, 2018; Sreelakshmi *et al.*, 2019; Lin, 2019; Deveci Topal *et al.*, 2021), les systèmes d'évaluation automatisée (*Automated assessment and feedback*) (Lee *et al.*, 2019; Sung *et al.*, 2021; Maestrales *et al.*, 2021), les outils de rédaction automatique comme les générateurs de textes [1] et les correcteurs grammaticaux automatisés [2], ou encore les applications assistées par l'IA tel que les outils de traduction [3] et de résolutions de problèmes mathématiques [4]. L'IA joue également un rôle important dans le soutien aux élèves en situation de handicap, grâce à des applications de diagnostic (dyslexie, TDAH, dysgraphie) (Barua *et al.*, 2022) et à des outils d'assistance (sous-titrage automatique, synthèse vocale, etc.). Pour les enseignants, nous pouvons trouver les détecteurs de plagiat [5], les plateformes de curation intelligente de contenus [6], les systèmes de monitoring en classe (Bosch & D'Mello, 2021), et les assistants à l'évaluation et à la correction. Enfin, du côté des institutions, nous allons voir des solutions d'IA pour optimiser la planification des cours (Kitto *et al.*, 2020), la détection des élèves à risque (Del Bonifro *et al.*, 2020), la gestion des admissions (Marcinkowski *et al.*, 2020), ou encore la surveillance automatisée des examens (e-proctoring) (Nigam *et al.*, 2021).

Avec l'apparition de l'IA générative, un nouveau paradigme est né, en particulier pour les systèmes de tutorat intelligents. En effet, grâce aux grands modèles de langue, il est maintenant possible de générer des contenus éducatifs dynamiques et pertinents en fonction du contexte (Maity & Deroy, 2024). Une des applications de l'IA générative concerne les systèmes de dialogue interactifs. Ces modèles permettent des échanges plus naturels et engageants entre l'apprenant et le système pour expliquer et répondre à ses questions. Malgré les avantages que l'intégration de l'IA générative apporte à l'éducation, en particulier dans les systèmes tutoriels intelligents, cette pratique soulève plusieurs défis majeurs. En effet, un des problèmes connus des LLMs est la possibilité de production d'informations erronées ou inappropriées, ainsi que l'impact de biais (stéréotypes, surreprésentation

1. https ://chatgpt.com/g/g-OolQ7FMzJ-ai-text-generator-gpt
2. https ://www.grammarly.com/
3. https ://www.deepl.com/translator
4. https ://photomath.com/
5. https ://plagiarismcheckerx.com, https ://www.turnitin.com/
6. https ://www.x5gon.org/

de certains points de vue, influence de la localisation des informations dans les documents...) durant l'apprentissage. Or, dans un contexte éducatif, pouvoir garantir la pertinence pédagogique du contenu généré est essentiel (Maity & Deroy, 2024).

Face aux défis soulevés par l'usage de l'IA générative dans les environnements éducatifs, il est nécessaire de concevoir des approches ciblées et fiables. C'est dans cette optique que nous proposons un système de question-réponse capable de répondre à des questions spécifiques en exploitant des supports privilégiés pour appuyer les réponses données.

### 2.2 Les systèmes RAG et l'éducation

Une des solutions qui a émergé pour atténuer ou pallier les failles des LLMs, en particulier le phénomène d'hallucination, est le *Retrieval Augmented Generation* (RAG) (Swacha & Gracel, 2025). Le RAG améliore les performances des LLMs en combinant une récupération d'informations pertinentes dans une base de données (ou base documentaire de référence) avec la génération de réponses en langage naturel, offrant ainsi des réponses plus précises et contextuellement adaptées.

Plusieurs applications existent dans la littérature qui intègrent le RAG afin d'interroger une source de connaissance externe : en domaine juridique (Wiratunga *et al.*, 2024; Cui *et al.*, 2024), pour les sciences de la santé et des sciences biomédicales (Li *et al.*, 2023; Lála *et al.*, 2023), la finance (Habib *et al.*, 2024), l'informatique (Dean *et al.*, 2023) ou encore pour plusieurs domaines (Forootani *et al.*, 2025). Malgré que ces applications ne sont pas spécifiquement faitent pour l'apprentissage, ce type de systèmes peut-être utilisé dans un contexte éducatif. Nous retrouvons bien évidemment plusieurs approches destinées à l'apprentissage (Jiang *et al.*, 2024; Abraham *et al.*, 2024; Levonian *et al.*, 2023; Al Ghadban *et al.*, 2023).

## 3 Approche et Méthodologie

Cette section présente de manière détaillée les fondements méthodologiques de notre étude, en articulant d'abord la description du jeu de données Science Questions (*SCIQ*) utilisé comme collection de tests, puis en exposant les différents systèmes et configurations expérimentales testés et expérimentés pour répondre aux problématiques posées, avant de détailler enfin la stratégie d'évaluation mise en œuvre pour mesurer rigoureusement la performance et la pertinence des approches proposées.

### 3.1 Jeu de données Science Questions "SCIQ"

Nous avons choisi d'utiliser l'ensemble de données SCIQ (SCIence Questions)[7] (Welbl *et al.*, 2017) pour évaluer notre système RAG dans un contexte éducatif. SCIQ a été obtenu par *crowdsourcing* et a été conçu pour l'entraînement et l'évaluation de modèles de question-réponse (Welbl *et al.*, 2017). Il se compose de plus de 13.000 questions à choix multiples, découpé en trois sous-ensembles (11.700 questions dans l'ensemble d'entraînement et 1000 questions dans chaque ensemble de validation et de test) couvrent un niveau allant de l'école primaire aux cours d'introduction à l'université, telles que la biologie, la physique, la chimie et les sciences de la Terre (Welbl *et al.*, 2017). Chaque élément

7. https ://huggingface.co/datasets/allenai/sciq

de cet ensemble contient une question, quatre réponses possibles, parmi lesquelles une seule est correcte, la majorité des questions sont accompagnées de paragraphes supplémentaires et d'éléments d'information à l'appui des bonnes réponses (Yu *et al.*, 2024). Ce jeu de données est en anglais.

| | |
|---|---|
| **Question** | *What type of organism is commonly used in preparation of foods such as cheese and yogurt ?* |
| **Réponses** | **A : mesophilic organisms**<br>B : protozoa<br>C : gymnosperms<br>D : viruses |
| **Support / Explication** | *Mesophiles grow best in moderate temperature, typically between 25 C and 40 C (77 F and 104 F). Mesophiles are often found living in or on the bodies of humans or other animals. The optimal growth temperature of many pathogenic mesophiles is 37 C (98 F), the normal human body temperature. Mesophilic organisms have important uses in food preparation, including cheese, yogurt, beer and wine.* |

TABLE 1 – Un exemple d'entrée de SCIQ, constitué d'une question, de quatre réponses (la bonne réponse est en gras), ainsi que du passage justifiant la bonne réponse.

Notre choix s'est porté sur ce jeu de données pour les raisons suivantes : d'un côté, il constitue un ensemble standard, reconnu pour l'évaluation de modèles de type Question-Réponse (*Question-Answering*, QA) dans un cadre éducatif (Liu *et al.*, 2024). En outre, les questions sont formulées de manière simple mais rigoureuse, proches de ce que l'on pourrait trouver dans un sujet d'examen réel (Welbl *et al.*, 2017), un exemple de question est présenté dans la table 1. Ces questions sont ouvertes et leur traitement nécessite d'identifier et de comprendre les connaissances scientifiques pertinentes, avant de suivre certains raisonnemment pour y répondre (Yu *et al.*, 2024). Ce type de questions, nous permet d'une part d'évaluer la capacité de récupération des connaissances scientifiques pertinentes, et d'autre part, d'évaluer la capacité de raisonner du modèle. Un autre avantage de SCIQ est le format des questions, à choix multiples, qui permet une évaluation automatique des réponses.

Ici, nous utilisons SCIQ pour comparer deux approches : un modèle génératif classique basé sur un LLM, et un système RAG dans lequel les réponses sont générées à partir d'un ensemble de documents récupérés automatiquement. Nous cherchons à évaluer non seulement la performance quantitative (précision des réponses), mais également la qualité des justifications fournies par le système.

## 3.2 Modèles et configurations expérimentés

### 3.2.1 Baseline : un modèle LLM seul

Comme référence de génération de réponse, nous avons choisi d'utiliser les grands modèles de langue (LLM) Llama 3[8] et Mistral 7b[9] seuls. Ils permettent de générer une réponse aux questions selon une stratégie "sans exemple" (*zero-shot*). Cette stratégie n'exploite pas de source externe d'informations

8. https://www.llama.com/models/llama-3/
9. https://mistral.ai/news/announcing-mistral-7b

mais se base uniquement sur celles acquises lors de l'entraînement du LLM. Le modèle reçoit la question telle quelle et génère directement une réponse. Plus précisément, nous utilisons les modèles *Llama3 8b* et *Mistral 7b*[10]. Un exemple d'invite est donné dans la figure 2 : Invite 1.

### 3.2.2 Stratégie RAG

L'architecture de notre approche RAG, présentée dans la figure 1, est constituée de :

- **module de récupération d'information (*Retriever*) :** ce module a pour rôle de rechercher et retourner les documents pertinents à partir d'une source externe, ici un corpus documentaire, afin de répondre à une question donnée ;
- **module de génération de la réponse (*Generator*) :** ce module génère une réponse en se basant sur les documents pertinents rassemblés précédemment ;
- **base de connaissance :** cette base a été construite à partir des explications des questions de l'ensemble de données SCIQ. Ces passages explicatifs (tableau 1) jouent le rôle d'un corpus de documents concis et fiables.

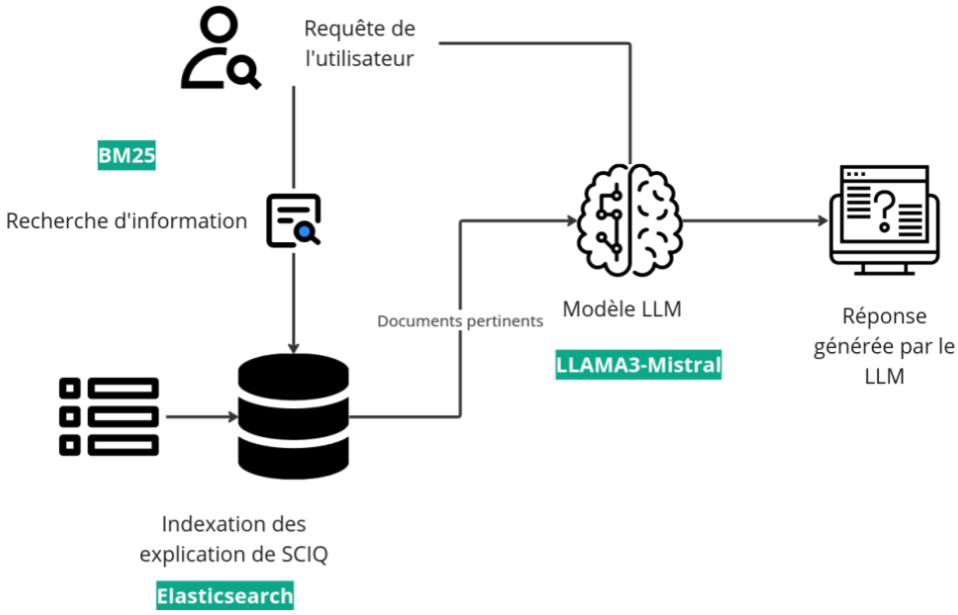



FIGURE 1 – Architecture de l'approche basée sur le RAG : des documents sont indexés puis récupérés par le moteur de recherche ElasticSearch en exploitant une approche BM25, les documents trouvés et la question sont transmises au LLM qui génère alors une réponse.

Nous commençons par construire une base de connaissance. Nous considérons les explications de chaque question de l'ensemble SCIQ comme autant d'unités documentaires à indexer dans le moteur de recherche ElasticSearch[11] avec un identifiant unique. ElasticSearch est une base de données NoSQL orientée documents, optimisée pour la recherche et l'analyse en temps réel, qui nous permet de créer différents types d'index et d'effectuer plusieurs types de recherche.

10. https://ollama.com/library/llama3 et https://ollama.com/library/mistral
11. https://www.elastic.co/fr/elasticsearch

Le système, le module RAG et le modèle de langue de référence (ici soit Llama 3, soit Mistral 7b), reçoivent, de la part de l'utilisateur, uniquement la question du jeu de données SCIQ en entrée. Aucune proposition de réponse (choix multiple) ni explication supplémentaire ne sont fournies au système dans l'invite (*prompt*). Le moteur de recherche documentaire ElasticSearch permet d'identifier des documents supposés pertinents pour une requête en utilisant une approche "sac de mots", avec une fonction de score de type BM25 (Robertson *et al.*, 1995). Les $k$ passages ayant les scores les plus élevés sont retenus. Après avoir testé plusieurs valeurs de k : 1,3,5,10 et 20, nous avons fixé $k = 10$. Ces documents sont ensuite transmis au module de génération de réponse. Les documents retournés par le Retriever sont injectés dans une invite structurée, utilisée pour générer la réponse finale via le LLM. Un exemple d'invite donné au LLM est donné dans la figure 2 : Invite 2. La formulation de l'invite encourage des réponses concises et adaptées au format du jeu de données.

**Invite 1 - stragégie de référence LLM seul**

*Answer the following question with a short and simple response (a few words only).*
*If you don't know the answer, say 'Not found'.*
*Question : "question SCIQ"*
*Answer :*

**Invite 2 - stratégie RAG**

*Using the context below, answer the question that follows with a short and simple response (a few words only). If the answer cannot be found in the context, say 'Not found'.*
*Context : "documents"*
*Question : "question SCIQ"*

FIGURE 2 – Exemple d'invites pour les stratégies LLM seul et RAG.

## 3.3 Protocole d'évaluation

**Evaluation de l'étape de recherche de documents** : nous utilisons l'outil standard trec_eval[12] qui permet d'évaluer la pertinence des documents récupérés en fonction d'une liste de référence (*gold standard*, construit ici en association à chaque question sa *bonne* explication telle que donnée dans SCIQ), à l'aide de métriques telles que le *Mean Average Precision (MAP)*, *Mean Reciprocal Rank (MRR)*, la *Precision@k*, le *Rappel@k* et le *Normalized Discounted Cumulative Gain (nDCG)*. Ces mesures permettent d'évaluer non seulement si les documents retournés sont pertinents, mais aussi leur position dans la liste de résultats, ce qui est crucial pour l'efficacité d'un système RAG, où seuls les premiers documents sont utilisés pour générer la réponse finale. Dans notre cas, nous nous concentrons sur le MRR et le Rappel (global et R@k), car notre référence de jugements de pertinence ne contient qu'un seul bon document pertinent par requête.

**Evaluation des réponses générées** : pour évaluer la qualité des réponses générées par le système, plusieurs métriques complémentaires ont été utilisées. Tout d'abord, une mesure de correspondance exacte (Exact Match) $EM1$ qui permet de vérifier si la réponse du système correspond mot à mot à la

12. https://trec.nist.gov/trec_eval

réponse attendue. Puis $EM2$ pour considérer le cas où la bonne réponse est une sous-chaîne de la réponse générée. Ensuite, la distance de Levenshtein (Navarro, 2001) $DL$ mesure la similarité entre la réponse générée et la réponse de référence en entier et $DL\_Part$ (partiel) en tenant compte des sous-chaînes les plus proches, offrant ainsi une évaluation plus tolérante aux variations de formulation. Des métriques classiques telles que la Précision $P$, le Rappel $R$ et le $F1 - score$ permettent quant à elles d'évaluer les réponses selon une échelle binaire, réponse correcte ou non. Pour aller au-delà de la simple comparaison lexicale, une mesure d'exactitude (*accurracy*) selon un score de similitude sémantique est également appliquée, à l'aide du modèle *paraphrase-MiniLM-L6-v2*[13] $SS1$ et le modèle *all-mpnet-base-v2*[14] $SS2$, afin d'évaluer si le sens général des réponses est proche de celui des références attendues. Enfin, un LLM peut lui-même être utilisé pour estimer la pertinence des réponses générées.

# 4 Résultats et discussion

## 4.1 Evaluation de la recherche de documents

Nous avons commencé par évaluer la capacité de notre système RAG à retourner les explications (documents) pertinentes pour les requêtes. Nous avons effectué ce type sur l'ensemble de données *train* de SCIQ (table 2).

| Modèle | MRR | Rappel | R@5 |
|---|---|---|---|
| RAG-Simple | 0,7247 | 0,8425 | 0,8175 |

TABLE 2 – Evaluation de la recherche de documents pertinents

— La valeur élevée du MRR (0,7247) indique que le premier document pertinent est généralement retrouvé parmi les premières positions du classement, ce qui témoigne de la capacité du système à retourner efficacement l'information pertinente dès les premiers résultats.
— Rappel global (0,8425) et R@5 (0,8175) sont tous les deux élevés : ce qui confirme que notre document pertinent est souvent dans le top 5.

## 4.2 Évaluation de la réponse générée

Nous avons ensuite évalué la qualité des réponses retournées par l'approche RAG en les comparant avec la référence LLM seul (tableau 3). La stratégie `rag-llama` (utilisant Llama3 avec RAG) obtient les meilleurs scores globaux sur la majorité des métriques (EM1, DL, SS1, SS2, Précision, F1) Les approches de base `bl-mist` et `bl-llama` avec les modèles llm Mistral et LLAMA3 utilisés sans RAG ont des performances globalement inférieures. Ces résultats confirment que notre solution basée sur le RAG peut effectivement améliorer la précision de la réponse. Nous notons aussi que le modèle Llama3 est globalement plus performant que Mistral 7b pour SCIQ.

13. https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2
14. https://huggingface.co/sentence-transformers/all-mpnet-base-v2

| Stratégie | EM1 | EM2 | DL | DL_Part | SS1 | SS2 | P | R | F1 | LLM |
|---|---|---|---|---|---|---|---|---|---|---|
| bl-mist | 0,27 | 0,51 | 0,29 | 0,61 | 0,41 | 0,36 | 0,22 | 0,54 | 0,31 | 0,82 |
| bl-llama | 0,39 | 0,46 | 0,42 | 0,59 | 0,51 | 0,47 | 0,45 | 0,47 | 0,46 | 0,75 |
| rag-llama | **0,46** | 0,56 | **0,50** | 0,68 | **0,59** | **0,55** | **0,49** | 0,56 | **0,52** | 0,79 |
| rag-mist | 0,38 | **0,74** | 0,41 | **0,82** | 0,54 | 0,47 | 0,15 | **0,78** | 0,25 | **0,94** |

TABLE 3 – Evaluation des réponses générées, avec RAG (rag) ou sans RAG (bl), avec Llama 3 (llama) ou Mistral 7b (mist).

## 4.3 Traçabilité des réponses : gestion de l'absence d'information

Nous nous plaçons maintenant dans un cadre de test où les questions posées ne bénéficient pas d'explication réponse explicite dans la base de connaissance (il n'y a pas de *document support*). Nous avons pris les questions de l'ensemble *test* du jeu de données SCIQ car les supports de ces derniers n'ont pas été indexés dans notre base de connaissance (cette dernière contient les supports de l'ensemble *train* seulement). Nous avons interrogé le modèle baseline **Llama3** avec ces question ainsi que notre système rag (représenté par bl-llama et rag-llama dans le tableau 4.3).

| Modèle | EM1 | EM2 | DL | DL_Part | SS1 | SS2 | P | R | F1 | LLM |
|---|---|---|---|---|---|---|---|---|---|---|
| bl-llama | **0,39** | **0,48** | **0,43** | **0,59** | **0,50** | **0,47** | **0,45** | **0,47** | **0,46** | **0,74** |
| rag-llama | 0,33 | 0,39 | 0,35 | 0,50 | 0,42 | 0,39 | 0,26 | 0,39 | 0,31 | 0,58 |

TABLE 4 – Evaluation de la réponse générée

— Nous remarquons une diminution très nette pour toutes les métriques du système RAG dans ce cas où les supports des questions ne sont pas présents dans la base de connaissance.
— Nous notons aussi que les métriques de notre baseline n'a pas été affecté par ce type test (comparé avec les résultats du tableau 3). La baseline a surpassé le RAG.

L'influence de l'absence de documents dans un système RAG se manifeste par des réponses plus prudentes ou absentes (dans ce cas le RAG nous a retourné plus de réponses de type "not found"). Dans le contexte éducatif, il est préférable de signaler un manque d'information plutôt que de générer une réponse incorrecte, grâce au RAG, l'hallucination du LLM est réduite.

## 4.4 Clarté des explications

La stratégie RAG permet de préciser le document support utilisé pour générer la réponse. De plus, nous pouvons indiquer au LLM du RAG d'expliquer la réponse et pourquoi avoir choisie celle-ci. Dans ce cas 4.4, le support utilisé par notre RAG est le même que celui qui devrait être utilisé, tel que mentionné dans l'ensemble de données. Nous avons retrouvé d'autres cas, où la réponse est correcte et le support utilisé par le module de génération est bien pertinent pour répondre à la question mais n'est pas celui attendu, comme dans le jeu de données.

| Question | Réponse | support | Explication |
|---|---|---|---|
| Transform faults are the site of massive what ? | earthquakes | ID_support :94<br>Transform faults are the site of massive earthquakes. | In the provided context, it is stated that “Transform faults are the site of massive earthquakes” (Paragraph 94). |
| In a glass of sweet tea the sugar is known as the solute and the water is known as what ? | solvent | ID_support :123<br>If you’re like Tanya in this picture, you prefer your iced tea sweetened with sugar. Sweetened iced tea is a solution in which solid sugar (the solute) is dissolved in cold liquid tea, which is mostly water (the solvent). When you add sugar to tea, particles of water pull apart particles of sugar. The particles of sugar spread throughout the tea, making all of it taste sweet. | In the context, the definition of iced tea (paragraph 123) states that solid sugar is dissolved in cold liquid tea. The soluble substance is the solute and the liquid it is dissolved in is the solvent. |

TABLE 5 – Exemples de réponses générées et d’explications associées.

# 5 Conclusion et perspectives

Plusieurs pistes d’amélioration restent ouvertes. Le jeu de données utilisé est bien connu et largement exploité dans la littérature, ce qui signifie que certains LLMs ont probablement été entraînés sur ce contenu. Cela peut expliquer les performances élevées observées pour les modèles de référence, LLM seuls. La recherche de documents pertinents reste peu performante en termes de classement des documents retrouvés, ce qui suggère la nécessité d’améliorer ce composant, notamment via des techniques de réordonnancement.

Cependant, l’intégration de l’approche RAG dans notre solution a permis d’améliorer la qualité des réponses générées, avec des gains notables notamment avec la stratégie `rag-llama`. Ces résultats confirment la pertinence du RAG dans un système de questions réponses, en particulier dans le contexte éducatif. En effet, dans ce cadre, ces améliorations sont significatives : une meilleure précision et un rappel élevé permettent de fournir aux apprenants des réponses plus fiables et complètes, tout en réduisant les risques liés aux hallucinations des modèles. Ainsi, les systèmes RAG peuvent contribuer à la mise en place d’outils d’assistance à l’apprentissage plus efficaces et mieux adaptés aux besoins spécifiques des élèves.

# Références

ABRAHAM S., EWARDS V. & TERENCE S. (2024). Interactive Video Virtual Assistant Framework with Retrieval Augmented Generation for E-Learning. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, p. 1192–1199. DOI : 10.1109/ICAAIC60222.2024.10575255.

AL GHADBAN Y., LU H. Y., ADAVI U., SHARMA A., GARA S., DAS N., KUMAR B., JOHN R., DEVARSETTY P. & HIRST J. E. (2023). Transforming Healthcare Education : Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation. DOI : 10.1101/2023.12.15.23300009.

BARUA P. D., VICNESH J., GURURAJAN R., OH S. L., PALMER E., AZIZAN M. M., KADRI N. A. & ACHARYA U. R. (2022). Artificial Intelligence Enabled Personalised Assistive Tools to Enhance Education of Children with Neurodevelopmental Disorders—A Review. *International Journal of Environmental Research and Public Health*, **19**(3), 1192. Number : 3 Publisher : Multidisciplinary Digital Publishing Institute, DOI : 10.3390/ijerph19031192.

BOSCH N. & D'MELLO S. K. (2021). Automatic Detection of Mind Wandering from Video in the Lab and in the Classroom. *IEEE Transactions on Affective Computing*, **12**(4), 974–988. DOI : 10.1109/TAFFC.2019.2908837.

CUI J., NING M., LI Z., CHEN B., YAN Y., LI H., LING B., TIAN Y. & YUAN L. (2024). Chatlaw : A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model. arXiv :2306.16092 [cs], DOI : 10.48550/arXiv.2306.16092.

DEAN M., BOND R. R., MCTEAR M. F. & MULVENNA M. D. (2023). ChatPapers : An AI Chatbot for Interacting with Academic Research. In *2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, p. 1–7, Letterkenny, Ireland : IEEE. DOI : 10.1109/AICS60730.2023.10470521.

DEL BONIFRO F., GABBRIELLI M., LISANTI G. & ZINGARO S. P. (2020). Student Dropout Prediction. In I. I. BITTENCOURT, M. CUKUROVA, K. MULDNER, R. LUCKIN & E. MILLÁN, Éds., *Artificial Intelligence in Education*, volume 12163, p. 129–140. Cham : Springer International Publishing. Series Title : Lecture Notes in Computer Science, DOI : 10.1007/978-3-030-52237-7_11.

DEVECI TOPAL A., DILEK EREN C. & KOLBURAN GEÇER A. (2021). Chatbot application in a 5th grade science course. *Education and Information Technologies*, **26**(5), 6241–6265. DOI : 10.1007/s10639-021-10627-8.

DOLENC K., ABERŠEK B. & KORDIGEL ABERŠEK M. (2015). ONLINE FUNCTIONAL LITERACY, INTELLIGENT TUTORING SYSTEMS AND SCIENCE EDUCATION. *Journal of Baltic Science Education*, **14**(2), 162–171. DOI : 10.33225/jbse/15.14.162.

FOROOTANI A., ALIABADI D. E. & THRAEN D. (2025). Bio-Eng-LMM AI Assist chatbot : A Comprehensive Tool for Research and Education. arXiv :2409.07110 [eess], DOI : 10.48550/arXiv.2409.07110.

HABIB M. A., AMIN S., OQBA M., JAIPAL S., KHAN M. J. & SAMAD A. (2024). TaxTajweez : A Large Language Model-based Chatbot for Income Tax Information In Pakistan Using Retrieval Augmented Generation (RAG). *The International FLAIRS Conference Proceedings*, **37**. DOI : 10.32473/flairs.37.1.135648.

HOLMES W. & TUOMI I. (2022). State of the art and practice in AI in education. *European Journal of Education*, **57**(4), 542–570. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/ejed.12533, DOI : 10.1111/ejed.12533.

JIANG Y., SHAO Y., MA D., SEMNANI S. J. & LAM M. S. (2024). Into the Unknown Unknowns : Engaged Human Learning through Participation in Language Model Agent Conversations. arXiv :2408.15232 [cs], DOI : 10.48550/arXiv.2408.15232.

KITTO K., SARATHY N., GROMOV A., LIU M., MUSIAL K. & BUCKINGHAM SHUM S. (2020). Towards skills-based curriculum analytics : can we automate the recognition of prior learning ? In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, p. 171–180, Frankfurt Germany : ACM. DOI : 10.1145/3375462.3375526.

LANE H. C., CAHILL C., FOUTZ S., AUERBACH D., NOREN D., LUSSENHOP C. & SWARTOUT W. (2013). The Effects of a Pedagogical Agent for Informal Science Education on Learner Behaviors and Self-efficacy. In D. HUTCHISON, T. KANADE, J. KITTLER, J. M. KLEINBERG, F. MATTERN, J. C. MITCHELL, M. NAOR, O. NIERSTRASZ, C. PANDU RANGAN, B. STEFFEN, M. SUDAN, D. TERZOPOULOS, D. TYGAR, M. Y. VARDI, G. WEIKUM, H. C. LANE, K. YACEF, J. MOSTOW & P. PAVLIK, Éds., *Artificial Intelligence in Education*, volume 7926, p. 309–318. Berlin, Heidelberg : Springer Berlin Heidelberg. Series Title : Lecture Notes in Computer Science, DOI : 10.1007/978-3-642-39112-5_32.

LEE H., PALLANT A., PRYPUTNIEWICZ S., LORD T., MULHOLLAND M. & LIU O. L. (2019). Automated text scoring and real-time adjustable feedback : Supporting revision of scientific arguments involving uncertainty. *Science Education*, **103**(3), 590–622. DOI : 10.1002/sce.21504.

LEVONIAN Z., LI C., ZHU W., GADE A., HENKEL O., POSTLE M.-E. & XING W. (2023). Retrieval-augmented Generation to Improve Math Question-Answering : Trade-offs Between Groundedness and Human Preference. arXiv :2310.03184 [cs], DOI : 10.48550/arXiv.2310.03184.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T. *et al.* (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, **33**, 9459–9474.

LI Y., LI Z., ZHANG K., DAN R., JIANG S. & ZHANG Y. (2023). ChatDoctor : A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus*. DOI : 10.7759/cureus.40895.

LIN Y.-H. (2019). A Supportive Information Assistant on Mobile Devices for Non-Technical Students Learning Programming.

LIU Y., CAO J., LIU C., DING K. & JIN L. (2024). Datasets for Large Language Models : A Comprehensive Survey. arXiv :2402.18041 [cs], DOI : 10.48550/arXiv.2402.18041.

LÁLA J., O'DONOGHUE O., SHTEDRITSKI A., COX S., RODRIQUES S. G. & WHITE A. D. (2023). PaperQA : Retrieval-Augmented Generative Agent for Scientific Research. arXiv :2312.07559 [cs], DOI : 10.48550/arXiv.2312.07559.

MAESTRALES S., ZHAI X., TOUITOU I., BAKER Q., SCHNEIDER B. & KRAJCIK J. (2021). Using Machine Learning to Score Multi-Dimensional Assessments of Chemistry and Physics. *Journal of Science Education and Technology*, **30**(2), 239–254. DOI : 10.1007/s10956-020-09895-9.

MAITY S. & DEROY A. (2024). Generative AI and Its Impact on Personalized Intelligent Tutoring Systems. arXiv :2410.10650 [cs], DOI : 10.48550/arXiv.2410.10650.

MARCINKOWSKI F., KIESLICH K., STARKE C. & LÜNICH M. (2020). Implications of AI (un-)fairness in higher education admissions : the effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, p. 122–130, Barcelona Spain : ACM. DOI : 10.1145/3351095.3372867.

MAYER R. E. (2019). Computer Games in Education. *Annual Review of Psychology*, **70**(Volume 70, 2019), 531–549. Publisher : Annual Reviews, DOI : 10.1146/annurev-psych-010418-102744.

MCLAREN B. M., ADAMS D. M., MAYER R. E. & FORLIZZI J. (2017). A Computer-Based Game that Promotes Mathematics Learning More than a Conventional Approach :. *International Journal of Game-Based Learning*, **7**(1), 36–56. DOI : 10.4018/IJGBL.2017010103.

NAVARRO G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, **33**(1), 31–88.

NIGAM A., PASRICHA R., SINGH T. & CHURI P. (2021). A Systematic Review on AI-based Proctoring Systems : Past, Present and Future. *Education and Information Technologies*, **26**(5), 6421–6445. DOI : 10.1007/s10639-021-10597-x.

PARONG J., MAYER R. E., FIORELLA L., MACNAMARA A., HOMER B. D. & PLASS J. L. (2017). Learning executive function skills by playing focused video games. *Contemporary Educational Psychology*, **51**, 141–151. DOI : 10.1016/j.cedpsych.2017.07.002.

PASCHOAL L. N., DE OLIVEIRA M. M. & CHICON P. M. M. (2018). A Chatterbot Sensitive to Student's Context to Help on Software Engineering Education. In *2018 XLIV Latin American Computer Conference (CLEI)*, p. 839–848, São Paulo, Brazil : IEEE. DOI : 10.1109/CLEI.2018.00105.

ROBERTSON S. E., WALKER S., JONES S., HANCOCK-BEAULIEU M. M., GATFORD M. *et al.* (1995). Okapi at trec-3. *Nist Special Publication Sp*, **109**, 109.

SREELAKSHMI A., ABHINAYA S., NAIR A. & JAYA NIRMALA S. (2019). A Question Answering and Quiz Generation Chatbot for Education. In *2019 Grace Hopper Celebration India (GHCI)*, p. 1–6, Bangalore, India : IEEE. DOI : 10.1109/GHCI47972.2019.9071832.

SUNG S. H., LI C., CHEN G., HUANG X., XIE C., MASSICOTTE J. & SHEN J. (2021). How Does Augmented Observation Facilitate Multimodal Representational Thinking ? Applying Deep Learning to Decode Complex Student Construct. *Journal of Science Education and Technology*, **30**(2), 210–226. DOI : 10.1007/s10956-020-09856-2.

SWACHA J. & GRACEL M. (2025). Retrieval-Augmented Generation (RAG) Chatbots for Education : A Survey of Applications. *Applied Sciences*, **15**(8), 4234. Number : 8 Publisher : Multidisciplinary Digital Publishing Institute, DOI : 10.3390/app15084234.

WARD W., COLE R., BOLAÑOS D., BUCHENROTH-MARTIN C., SVIRSKY E. & WESTON T. (2013). My science tutor : A conversational multimedia virtual tutor. *Journal of Educational Psychology*, **105**(4), 1115–1125. DOI : 10.1037/a0031589.

WELBL J., LIU N. F. & GARDNER M. (2017). Crowdsourcing Multiple Choice Science Questions. arXiv :1707.06209 [cs], DOI : 10.48550/arXiv.1707.06209.

WIRATUNGA N., ABEYRATNE R., JAYAWARDENA L., MARTIN K., MASSIE S., NKISI-ORJI I., WEERASINGHE R., LIRET A. & FLEISCH B. (2024). CBR-RAG : Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering. arXiv :2404.04302 [cs], DOI : 10.48550/arXiv.2404.04302.

YU H.-C., SHIH Y.-A., LAW K.-M., HSIEH K.-Y., CHENG Y.-C., HO H.-C., LIN Z.-A., HSU W.-C. & FAN Y.-C. (2024). Enhancing Distractor Generation for Multiple-Choice Questions with Retrieval Augmented Pretraining and Knowledge Graph Integration. arXiv :2406.13578 [cs], DOI : 10.48550/arXiv.2406.13578.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# *InitIAtion* : développer l'agentivité numérique au collégial à l'ère de l'intelligence artificielle générative

Fanny Joussemet[1, 2, 3]
(1) Cégep de Saint-Laurent, 625 av. Sainte-Croix, Saint-Laurent, QC H4L 3X6, Canada
(2) AQPC — Association québécoise de pédagogie collégiale, 999 Av. Émile-Journault, Montréal, Canada
(3) OBVIA — Observatoire sur les impacts sociétaux de l'IA et du numérique, Pavillon Charles-De Koninck, local 2489, 1030, avenue des Sciences-Humaines, Université Laval Québec QC G1V 0A6, Canada
fjoussemet@cegepsl.qc.ca

## RÉSUMÉ

Cet article présente la genèse, le cadre conceptuel et les principes pédagogiques de la trousse *InitIAtion*, conçue pour accompagner les étudiantes et étudiants du collégial dans un usage critique, responsable et créatif de l'intelligence artificielle générative (IAg). Issu de consultations et d'expérimentations menées au Cégep de Saint-Laurent, ce projet répond à un besoin identifié de formation structurée, face à une adoption rapide de l'IAg dans les pratiques étudiantes. Appuyé sur le référentiel international de l'UNESCO et sur les recommandations du Conseil supérieur de l'éducation du Québec, *InitIAtion* se structure autour des compétences associées au métier d'étudiant. Proposée sous forme modulaire, adaptable à divers contextes disciplinaires, la trousse de formation vise à développer l'agentivité numérique, la pensée critique et l'autonomie des étudiantes et des étudiants. L'article discute également des modalités d'implantation, dans une logique de mutualisation interordres, et appelle à une appropriation contextualisée de l'outil par les établissements d'enseignement supérieur québécois.

## ABSTRACT

***InitIAtion* : Developing Digital Agency in College in the Age of Generative Artificial Intelligence**

This article presents the origins, conceptual framework and pedagogical principles of the *InitIAtion* toolkit, designed to support college students in a critical, responsible and creative use of generative artificial intelligence (GenAI). Based on consultations and experiments conducted at the Cégep de Saint-Laurent, this project responds to an identified need for structured education in response to the rapid adoption of GenAI in student practices. Grounded in UNESCO's international reference framework and the recommendations of the Conseil supérieur de l'éducation du Québec, *InitIAtion* is structured around the competencies associated with the student role. Offered in modular form, adaptable to various disciplinary contexts, the toolkit aims to develop students' digital agency, critical thinking and autonomy. The article also discusses the modalities of implementation, in a logic of inter-order mutualization, and calls for a contextualized appropriation of the tool by Quebec higher education institutions.

MOTS-CLÉS : intelligence artificiell générative, littératie numérique, agentivité numérique, enseignement supérieur, cégep, éducation.

KEYWORDS: Generative Artificial Intelligence, Digital Literacy, Digital Agency, Higher Education, cégep, Education.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 42–62.

# 1 Introduction

À l'automne 2023 et à l'hiver 2024, le Cégep de Saint-Laurent a mené le projet *IA & Réussite* pour mieux comprendre les usages, perceptions et besoins liés à lintelligence artificielle générative (IAg) en enseignement collégial québécois [1], notamment via les grands modèles de langage (GML, en anglais LLM *Large Language Models*). Bien que non récente, l'IAg — définie comme la capacité de produire de nouveaux contenus à partir de données massives (Centre canadien pour la cybersécurité, 2023) — a connu une diffusion accélérée dès la fin de 2022. L'ampleur de cette adoption, tant personnelle que professionnelle, ainsi que les changements dans les représentations sociales qu'elle a entraînés justifient son statut d'innovation technologique significative (Verchère, 2024). Les constats tirés de ce projet ont permis d'identifier des enjeux majeurs, à la base de la conception de la trousse *InitIAtion*.

En effet, dans un contexte de reconfiguration des repères pédagogiques [2], une première documentation des usages a révélé, à l'automne 2023, que 27% des 928 étudiantes et étudiants sondés au cégep de Saint-Laurent utilisaient des GML comme ChatGPT à des fins académiques, tous programmes confondus (Joussemet *et al.*, 2024c). Bien que ce taux ait été inférieur aux perceptions initiales des responsables du projet, ces usages portaient déjà sur des tâches variées (résumés, idées, correction linguistique, recherche d'information), moins d'un an après la mise en ligne de ChatGPT (Peters, 2023). Le contexte, marqué par des réactions fortes et une « chasse au ChatGPT » [3], a pu influencer la sincérité et la précision des réponses.

L'appropriation transversale de l'IAg restait variable selon les domaines de formation, mais la familiarité avec l'outil demeurait croissante malgré des malentendus sur ses fonctions réelles, notamment en matière de recherche d'information et de citation. Les étudiants y voyaient un gain de temps et un soutien, tout en s'inquiétant de la fiabilité, du plagiat involontaire et de la perte de compétences (Joussemet *et al.*, 2024c). Ces constats rejoignaient ceux d'études pointant les bénéfices des usages de l'IAg sur la personnalisation pédagogique, l'idéation, et le soutien rédactionnel, notamment pour les allophones (Abdelghani *et al.*, 2023; Chan & Hu, 2023), tout en rappelant les risques de passivité, d'erreurs non détectées et de dépendance (Abdelghani *et al.*, 2023; Chan & Lee, 2023).

Au cégep de Saint-Laurent, moins d'une session plus tard, un second sondage révélait une évolution notable : 71,1% des étudiantes et étudiants en sciences humaines (n = 45) déclaraient avoir utilisé l'IAg, contre 31,8% une session plus tôt (n = 151) (Joussemet *et al.*, 2024c). Ce basculement rapide, bien que fondé sur un petit échantillon, montrait une appropriation extrêmement rapide. Les étudiantes et les étudiants exprimaient des attentes élevées (gain de temps, soutien, idées), mais aussi des incertitudes et un besoin d'encadrement éthique et méthodologique : « Le cégep n'a pas beaucoup de pouvoir concernant l'utilisation de ChatGPT. [...] Il faut bien l'encadrer pour bien l'utiliser » (étudiant en génie électrique et informatique, cité dans (Joussemet & Meurs, 2024) (diapositive 30).

En parallèle, des groupes de discussion avec 18 enseignantes et enseignants issus des trois secteurs de la formation collégiale (générale, préuniversitaire et technique) mettaient en lumière des préoccupations convergentes (Joussemet *et al.*, 2023) : besoins de formation, crainte du plagiat, impact sur les apprentissages, manque de temps pour s'adapter et ajuster ses évaluations. Ces constats rejoignaient

1. Les cégeps (Collèges d'enseignement général et professionnel) forment un réseau d'établissements publics d'enseignement supérieur créés en 1967 au Québec. Ils accueillent les étudiantes et les étudiants après l'école secondaire (équivalente au lycée) et offrent à la fois des formations préuniversitaires (menant à l'université) et techniques (orientées vers le marché du travail) - Qu'est-ce que le cégep?
2. ChatGPT: des profs de cégep et des chargés de cours tirent la sonnette d'alarme, 15/05/2023 à Journal de Québec
3. La chasse au ChatGPT est ouverte, 20/01/2023 à La Presse.

ceux de l'Association québécoise de pédagogie collégiale (AQPC) sur l'accroissement des besoins en formation continue et le décalage croissant entre usages étudiants de la technologie et capacité institutionnelle d'encadrement (Association québécoise de pédagogie collégiale, 2023). Cette diversité de positions soulignait l'importance de construire une approche contextualisée et prudente, fondée sur l'esprit critique, l'éthique et l'expérimentation encadrée.

Ces observations ont conduit à une question centrale : comment accompagner les étudiantes et les étudiants dans un usage critique et responsable de l'IAg, adapté à l'enseignement collégial ? D'autant plus que le vide laissé par les établissements commençait à être comblé par des formations privées, souvent offertes par les entreprises propriétaires de GML, soulevant des enjeux d'indépendance et de finalité. Celles-ci privilégiant des « compétences techniques nécessaires à l'exploitation de plateformes d'IA orientées vers le profit » et négligeant les enjeux critiques liés à l'apprentissage et à la citoyenneté (Fengchun & Shiohira, 2024, p. 12, traduction personnelle)

Pour y répondre, quatre priorités ont été formulées par l'équipe responsable du projet : baliser, former, outiller, expérimenter (Gosselin *et al.*, 2024). Ces axes s'inscrivent dans la continuité des recommandations formulées par l'AQPC dans son mémoire au Conseil supérieur de l'éducation (CSÉ), qui soulignent l'importance de développer la littératie numérique, d'assurer la formation continue, de garantir l'équité d'accès aux ressources et de soutenir l'innovation pédagogique (Association québécoise de pédagogie collégiale, 2023).

C'est dans ce contexte d'intérêt et de besoins pédagogiques encore non comblés que la trousse *InitIAtion* a vu le jour. Conçue pour former les étudiantes et les étudiants dès leur entrée dans l'enseignement supérieur, elle vise à développer leur agentivité numérique, leur autonomie et leur pensée critique à propos de l'IAg. La conception « clé en main » de la trousse se veut aussi être une réponse aux contraintes de temps du corps enseignant face à des contenus complexes et évolutifs (Conseil supérieur de l'éducation & Commission de l'éthique en science et en technologie, 2024; Joussemet *et al.*, 2023).

Cet article retrace la trajectoire du projet *InitIAtion*, de son élaboration à partir d'expérimentations locales à son développement interordres avec le Pôle interordres de Montréal (PIM)[4], en précisant son cadre conceptuel, sa structure modulaire et ses perspectives de déploiement dans les établissements d'enseignement supérieur.

# 2 De l'expérimentation locale à la mutualisation interordres : la trajectoire d'*InitIAtion*

## 2.1 Présentation exploratoire de la trousse au cégep de Saint-Laurent

À la suite des consultations conduites à l'hiver 2024, une version préliminaire de la trousse *InitIAtion* a été conçue. Elle comprenait un ensemble de ressources complémentaires, structurées afin den favoriser l'appropriation pédagogique. Parmi celles-ci figurait un diaporama destiné au corps professoral, abordant les notions fondamentales relatives à l'IA, à l'IAg, aux GML, ainsi quaux enjeux de qualité de l'information et d'intégrité intellectuelle. Ce support détaillait les intentions pédagogiques et les objectifs des activités proposées. Un diaporama simplifié, conçu spécifiquement pour les étudiantes

4. https://pim.quebec/

et étudiants, accompagnait ce matériel en vue d'un usage en classe (Joussemet *et al.*, 2024b). Des éléments de référence au plan-cadre du cours étaient également fournis, afin d'assurer une intégration cohérente des contenus. Un exercice de discussion portant sur les balises d'utilisation de l'IAg, à réaliser en classe, accompagnait ces supports, de même qu'un questionnaire interactif visant à susciter une réflexion collective sur les usages de cette technologie. Enfin, une fiche synthèse, élaborée en collaboration avec le personnel bibliothécaire, proposait une sélection d'outils de recherche documentaire spécialisés, jugés plus pertinents que les modèles génératifs généralistes, tels que ChatGPT, pour répondre aux besoins académiques (Joussemet *et al.*, 2024a).

Le matériel a dabord été présenté à l'équipe enseignante du cours Réussir au collégial (360-RÉU), propre au Cégep de Saint-Laurent (Cégep de Saint-Laurent, 2022a). Ce cours de quinze heures, instauré à l'automne 2022, vise à faciliter la transition entre le secondaire et le collégial, en développant chez les étudiantes et étudiants des stratégies d'apprentissage, de gestion du temps, de recherche documentaire et d'appropriation des outils numériques (Dufour & Tardif, 2023). Rattaché aux activités de mise à niveau — définies comme favorisant l'acquisition de compétences essentielles à la poursuite des études (Ministère de l'Éducation et de l'Enseignement supérieur, 2018, p. 1) —, il a pour objectif (compétence 1006) l'utilisation de stratégies d'apprentissage efficaces (Ministère de l'Éducation et de l'Enseignement supérieur, 2018). Offert dans tous les programmes, RÉU constitue un espace propice à l'introduction de thématiques sur le « métier étudiant » et à la collaboration entre personnes enseignantes et services de soutien. Son positionnement stratégique (première session) et ses objectifs, dont « l'utilisation appropriée des technologies » comme critère de performance (Cégep de Saint-Laurent, 2022a, p. 2), en faisaient un terrain d'implantation pertinent pour *InitIAtion*.

La présentation visait à faire découvrir le matériel, recueillir des rétroactions et proposer des modalités de déploiement souples. Sans obligation institutionnelle, chaque enseignante ou enseignant pouvait intégrer tout ou partie de la trousse selon ses objectifs, sa familiarité avec le sujet ou les besoins de ses étudiantes et étudiants. Cette flexibilité respectait la diversité des contextes pédagogiques et favorisait une appropriation progressive. Pour faciliter l'implantation, il a été proposé qu'une personne-ressource anime la séance, en présence de la personne titulaire et du groupe, dès l'automne 2024.

## 2.2 2.2 Difficultés rencontrées et ajustements envisagés

L'implantation de la trousse dans les cours RÉU n'a pas suscité une adoption généralisée à l'automne 2024. Une consultation menée à l'hiver 2025 auprès des responsables des cours RÉU volontaires (Joussemet, 2025), issus de disciplines variées (sciences humaines, danse, architecture), a permis den cerner plusieurs freins. La trousse a été jugée trop volumineuse pour les formats courts comme RÉU, dispensé à raison d'une heure par semaine, et difficile à intégrer dans des périodes de 50 minutes. Les présentations ont été perçues comme « chargées », et les contraintes matérielles (absence d'ordinateurs, studios non équipés) limitaient les possibilités d'expérimenter les outils en classe. Dans plusieurs cas, leur approche de l'IAg est donc restée théorique. Une complexité à intégrer la trousse dans les séquences pédagogiques, déjà centrées sur les compétences élémentaires liées à la rédaction, à la recherche documentaire ou à la gestion des études a également été nommée par le corps enseignant.

D'autres difficultés, liées à la nature même du cours RÉU (Dufour, 2023; Régis, 2024), sont également venues limiter l'adoption de la trousse. Ce cours a en effet pris des formes variables selon les programmes dans lesquels il a été implanté, ce qui réduisait la possibilité d'y intégrer un contenu

uniformisé. Enfin, il convient de rappeler que le sujet de l'IAg suscite des réactions diverses dans le milieu de l'enseignement supérieur, allant de l'enthousiasme à la méfiance[5] [6]. Lors de la consultation (Joussemet, 2025), certains et certaines ont exprimé la crainte que l'introduction du sujet n'encourage « la triche » et préféraient ne pas aborder le sujet. Ainsi, la liberté laissée à chacun daborder — ou non — ces enjeux, en fonction de son aisance avec la technologie ou de ses propres balises éthiques, a également contribué à une appropriation inégale du matériel.

Ces observations ont conduit à repenser les modalités de déploiement de la trousse et à adapter son contenu, afin de répondre plus finement aux besoins exprimés sur le terrain.

## 2.3 Évolution des usages étudiants et nécessité d'une réponse élargie

Parallèlement, les données confirmaient que l'utilisation de l'IAg devenait une pratique courante : un sondage de KPMG indiquait que six étudiants sur dix l'utilisaient pour accomplir des tâches scolaires (KPMG, 2024). Le Sondage sur la population étudiante des cégeps (SPEC), réalisé en 2024 (n = 30 202), révélait que 59,9% des répondants (n = 23 550) avaient expérimenté un outil d'IAg, un taux atteignant 63,3% chez les moins de 18 ans, et 55,7% (n = 23 581) exprimaient un enthousiasme modéré à élevé envers ces technologies (Abran, 2024).

Quelques mois plus tard, les constats formulés par les enseignants et enseignantes lors de la consultation de l'hiver 2025 (Joussemet, 2025) sont venus donner un relief qualitatif à ces tendances chiffrées. Ces observations confirmaient non seulement que les usages étaient déjà bien implantés, mais validaient également les préoccupations exprimées sur le manque de recul critique, la méconnaissance du fonctionnement des outils et l'utilisation souvent superficielle de l'IAg. Ensemble, ces éléments ont renforcé la nécessité d'une réponse éducative mieux arrimée aux réalités du terrain.

Et pourtant, jusquà aujourd'hui, aucun cégep na instauré de formation systématique sur les usages responsables de ces outils, laissant place à des pratiques autonomes, peu balisées sur les plans éthique, critique et réflexif. Ce vide a commencé à se combler tout récemment, au printemps 2025, avec l'annonce du Cégep de Sainte-Foy d'une formation obligatoire de 30 minutes sur l'IAg destinée à toute la communauté étudiante dès l'automne, en réponse à une forte hausse des cas de plagiat[7].

Le besoin d'un encadrement structuré est confirmé par la recherche sur les effets de l'IAg sur l'apprentissage. (Abdelghani *et al.*, 2023) montrent que, si les GML peuvent stimuler l'engagement lorsqu'ils sont intégrés à des activités pédagogiques ciblées, leur usage autonome favorise des comportements passifs, une surestimation de soi et une baisse de la pensée critique. La conception de ces modèles, qui livrent des réponses rapides et affirmatives sans signaler les incertitudes, nuit au développement des compétences métacognitives, comme lautoévaluation ou l'ajustement de la compréhension. En l'absence de dispositifs éducatifs explicites, une véritable littératie critique ne peut émerger. (Jabagi & Croteau, 2025) ajoutent que « sans une formation et un positionnement adéquats, l'accès facile à ChatGPT peut conduire à une dépendance aveugle ou excessive qui peut entraver l'autonomie des personnes apprenantes ». Ces conclusions exigent une réflexion approfondie pour que l'IAg soutienne, plutôt qu'affaiblisse, l'autonomie et la pensée critique.

5. Des enseignants se tournent vers l'intelligence artificielle pour alléger leur tâche, 27/08/2023 à L'actualité
6. Face au plagiat, des professeurs appellent à un moratoire sur le développement de l'IA, 15/05/2023 à radio-canada
7. Offensive pour contrer le plagiat avec ChatGPT au cégep, 11/04/2025 à Journal de Québec.

## 2.4 Vers une mutualisation interordres : projet soutenu par le PIM

Face aux constats convergents issus des milieux scientifiques et institutionnels, la nécessité d'une formation structurée pour les étudiantes et étudiants s'est imposée. Ce besoin a trouvé un écho dans un appel à projets du PIM[8], qui visait à mutualiser les ressources et soutenir le développement d'outils pédagogiques en IAg, transférables à l'ensemble du réseau collégial et universitaire québécois. C'est dans ce cadre qu'a été retenue une proposition portée par une professeure de sociologie du cégep de Saint-Laurent — coresponsable du projet IA & Réussite et conceptrice du cours transdisciplinaire IA au quotidien (Cégep de Saint-Laurent, 2022b) — en partenariat avec le Carrefour d'innovation et de pédagogie universitaire de l'UQAM. Le projet visait à développer, médiatiser et diffuser la trousse *InitIAtion*, conçue spécifiquement pour accompagner les étudiantes et étudiants dans l'acquisition progressive d'une littératie critique, responsable et créative de l'IAg, adaptée à la diversité des contextes et des parcours de formation.

À notre connaissance, *InitIAtion* constitue un projet unique dans le réseau de l'enseignement supérieur québécois, en raison de son ampleur et de sa profondeur pédagogique : loin d'une formation ponctuelle ou superficielle, elle propose un parcours structuré, modulaire et progressif, pouvant accompagner les étudiantes et étudiants tout au long de leur cheminement collégial. Ce format vise à assurer une continuité éducative en littératie de l'IAg, alignée sur l'évolution des besoins de formation et l'acquisition graduelle de compétences critiques, éthiques et méthodologiques.

Cette approche prudente et graduelle répond à une préoccupation centrale dans la recherche : un usage trop précoce des GML pourrait compromettre le développement des compétences métacognitives des étudiantes et étudiants novices. (Abdelghani *et al.*, 2023, p. 4, traduction personnelle) rappellent qu'un recours hâtif à l'IAg « peut réduire l'engagement cognitif actif », en limitant l'autoévaluation, la confrontation à lerreur et la consolidation des savoirs. À l'entrée au collégial, il importe donc de privilégier l'appui sur leurs connaissances, l'acceptation de l'incertitude et la construction de leur confiance. La trousse *InitIAtion* s'inscrit ainsi dans une pédagogie de l'agentivité progressive, où l'IAg est un levier à maîtriser avec discernement, non une réponse immédiate.

Le parcours proposé repose sur quatre étapes complémentaires, qui jalonnent la progression attendue : comprendre (connaître le fonctionnement, les possibilités et les limites des outils), tester (s'exercer à l'art de la requête pour formuler des questionnements adaptés), questionner (évaluer de manière critique l'information produite par l'IAg), et utiliser (faire un usage éthique, responsable et créatif de ces outils).

Le projet redéfini répond à plusieurs constats, issus tant dobservations locales que d'analyses à l'échelle du réseau de l'enseignement supérieur. Il comble l'absence d'un cadre clair sur l'usage académique de l'IAg (ORES, 2024) et vise à corriger les inégalités d'accès à la formation, observées selon les disciplines, programmes ou profils étudiants (Conseil supérieur de l'éducation & Commission de l'éthique en science et en technologie, 2024; ORES, 2024). Car, sans action, cette fracture numérique pourrait accentuer les écarts de compétence et nuire à une préparation équitable au marché du travail, ou plus largement à la vie en société. (Jabagi & Croteau, 2025) rappellent qu'une formation accessible à tous « réduit les éventuelles fractures numériques fondées sur les compétences » (p. 13) et renforce l'appartenance des étudiantes et étudiants.

La trousse *InitIAtion* s'inscrit dans une conception de la littératie en IA, entendue comme l'« ensemble de compétences qui permet aux individus d'évaluer de manière critique les technologies d'IA, de

8. Qu'est-ce que le PIM ?

communiquer et de collaborer efficacement avec l'IA, et d'utiliser l'IA comme un outil en ligne, à la maison et au travail » (Groupe de travail sur la sensibilisation du public, 2022, p. 6). Cette approche est indissociable d'une reconnaissance des exigences propres au métier étudiant, compris comme un processus d'acquisition progressive de compétences transversales essentielles à la réussite au collégial : autonomie, rigueur, capacité danalyse et de réflexion critique.

Afin de concrétiser cette double visée — littératie de l'IAg et développement du métier étudiant — son élaboration intègre de manière structurante un cadre conceptuel combinant des repères internationaux et nationaux. Ce cadre, fondé sur les travaux de l'UNESCO et du CSÉ du Québec, guide la définition des objectifs, l'organisation des contenus et les modalités d'apprentissage.

# 3 Un cadre conceptuel à la croisée des référentiels internationaux et des valeurs éducatives québécoises

La trousse *InitIAtion* s'inscrit dans une approche pédagogique rigoureuse, fondée sur un cadre conceptuel qui articule des repères internationaux et les spécificités du contexte collégial québécois. Ce cadre repose principalement sur deux références : le Référentiel de compétences en IA pour les apprenants publié par l'UNESCO (Fengchun & Shiohira, 2024), et les recommandations du Conseil supérieur de l'éducation du Québec (Conseil supérieur de l'éducation & Commission de l'éthique en science et en technologie, 2024). Ensemble, ces orientations ont permis de structurer les contenus et les finalités de la trousse de manière à répondre aux enjeux éducatifs, éthiques et méthodologiques liés à lessor de l'IAg dans l'enseignement supérieur.

## 3.1 Le référentiel de l'UNESCO : une littératie en IA centrée sur l'agentivité des apprenants

L'UNESCO souligne la nécessité pour les apprenantes et apprenants d'atteindre un niveau suffisant d'alphabétisation en IA, incluant la compréhension de son fonctionnement général, de ses impacts sur la vie quotidienne, ainsi que des compétences adaptées à une utilisation pertinente et créative, notamment des outils génératifs (Fengchun & Wayne, 2024).

Pour guider ces formations, elle propose un référentiel (Fengchun & Shiohira, 2024) structuré, précisant les axes de compétences nécessaires à une appropriation complète et responsable de l'IA. Celui-ci repose sur quatre aspects complémentaires. Le premier met l'accent sur une perspective centrée sur l'humain, visant une compréhension critique des bénéfices et risques de l'IA, et du principe de proportionnalité entre les outils utilisés, les besoins humains et les enjeux environnementaux. Le second concerne l'éthique de l'IA, englobant les compétences sociales et morales requises pour naviguer dans un ensemble croissant de principes régissant un usage responsable tout au long du cycle de vie de ces technologies. Le troisième porte sur les techniques et applications de l'IA, couvrant les connaissances conceptuelles sur leur fonctionnement et les compétences pratiques nécessaires à leur utilisation dans des tâches authentiques. Enfin, le quatrième traite de la conception de systèmes dIA.

Chaque aspect est associé à trois niveaux progressifs : comprendre, appliquer, créer. Le niveau « comprendre » renvoie aux notions fondamentales de l'IA, ses modes de fonctionnement, ses enjeux sociaux et techniques. Le niveau « appliquer » engage les apprenantes et apprenants à

utiliser concrètement des outils dans des contextes réels, en mobilisant leurs connaissances avec discernement. Le niveau « créer » favorise des usages inventifs, critiques et responsables, où l'individu est capable d'adapter ou concevoir des solutions selon ses besoins. Cette progression assure une montée cohérente en compétence, de l'acquisition des concepts jusqu'à leur mobilisation autonome (Fengchun & Shiohira, 2024).

C'est dans cette logique que s'inscrit la trousse *InitIAtion*, qui accompagne les étudiantes et étudiants à travers ces trois niveaux — comprendre, appliquer, créer. En privilégiant les trois premiers aspects du référentiel de l'UNESCO — perspective humaine, éthique, techniques et applications de l'IAg — *InitIAtion* ancre l'apprentissage dans des situations académiques concrètes, adaptées aux réalités de l'enseignement collégial.

Ce cadre constitue la fondation de la trousse, qui repose sur une progression en quatre étapes — comprendre, tester, questionner, utiliser — visant le développement d'une agentivité numérique éclairée. Cette progression s'inspire aussi de la taxonomie de Bloom révisée[9], mettant l'accent sur les niveaux cognitifs supérieurs : analyser, évaluer, créer. *InitIAtion* mobilise ainsi plusieurs dimensions du savoir (Wilson, 2016) : les connaissances factuelles, pour comprendre les notions de base liées à l'IA, l'IAg et aux GML ; les connaissances conceptuelles, pour situer ces notions dans des cadres théoriques (biais, limites, implications sociales) ; les connaissances procédurales, mobilisées pour expérimenter les outils, formuler des requêtes efficaces et évaluer la qualité des contenus générés. La trousse favorise aussi le développement de connaissances métacognitives, en encourageant les apprenants et les apprenantes à une réflexion sur leurs usages, sur les conditions d'un recours pertinent à l'IAg, et sur les stratégies pour exercer un jugement critique et autonome dans divers contextes. *InitIAtion* dépasse ainsi la transmission de savoirs, en les engageant dans une appropriation active, fondée sur la compréhension des fonctions et limites des outils, l'élaboration de jugements critiques, et leur intégration raisonnée dans les pratiques académiques.

Cette approche s'inscrit pleinement dans les finalités de la formation collégiale québécoise, visant à former des citoyennes et citoyens capables de penser de façon critique, de résoudre des problèmes complexes, et de s'adapter aux évolutions rapides de la société (Gosselin, 2021).

## 3.2 Les recommandations du CSÉ : une appropriation de l'IAg fondée sur l'éthique, l'autonomie et l'intégrité

Le projet *InitIAtion* s'appuie également sur les recommandations formulées par le CSÉ et la CEST du Québec (2024). Le rapport examine les enjeux éthiques liés à l'IAg dans l'enseignement supérieur. À travers une méthodologie rigoureuse, comprenant une revue de la littérature scientifique, des consultations d'experts, ainsi qu'une analyse des pratiques actuelles dans les établissements, ce document formule vingt recommandations destinées à garantir une utilisation judicieuse de l'IAg dans les collèges et les universités, dans l'hypothèse où son usage deviendrait progressivement normalisé.

Dans le contexte québécois, où les collèges font partie intégrante de l'enseignement supérieur, ces recommandations reconnaissent l'importance fondamentale de l'autonomie institutionnelle, de la liberté académique et du respect de l'autonomie professionnelle des enseignantes et enseignants. Le CSÉ adopte ainsi une approche prudente et nuancée, invitant les établissements à réfléchir à la place qu'ils souhaitent accorder à l'IAg, sans prescrire une intégration systématique ou uniforme.

9. Taxonomie de Bloom révisée (domaine cognitif)

Cinq éléments prioritaires sont mis en avant par le rapport comme devant guider toute réflexion sur l'usage de l'IAg en milieu éducatif : l'alignement pédagogique des pratiques, le respect de l'intégrité intellectuelle, la formation continue à la compétence numérique, la qualité de l'information produite par les IAg, ainsi que la prise en compte des enjeux éthiques, tels que la protection de la vie privée ou les impacts environnementaux. À ces principes, le rapport du (Conseil supérieur de l'éducation & Commission de l'éthique en science et en technologie, 2024) ajoute l'importance de développer chez les étudiantes et étudiants des connaissances fondamentales sur le fonctionnement des IAg, leur capacité à évaluer la qualité des données sur lesquelles ces outils reposent, et à comprendre les logiques probabilistes qui sous-tendent leurs réponses. Il insiste également sur la nécessité d'aborder des notions clés telles que les biais algorithmiques, la transparence des modèles, et les enjeux tels que l'anthropomorphisme. Ces contenus sont jugés essentiels pour assurer une compréhension critique et favoriser des usages éclairés de l'IAg dans l'enseignement supérieur. Ces principes ont directement inspiré la conception d'*InitIAtion*, qui a été pensée comme un outil permettant aux milieux éducatifs de structurer une réponse adaptée à ces défis, tout en respectant leurs spécificités locales.

Loin de proposer un modèle prescriptif ou une voie unique d'utilisation de l'IAg, *InitIAtion* s'inscrit dans la logique préconisée par le CSÉ, en offrant un cadre pédagogique flexible, susceptible d'être adapté, enrichi ou modulé par les équipes enseignantes. Elle vise ainsi à accompagner, de manière lucide et contextualisée, les usages déjà présents dans les pratiques étudiantes, en favorisant une appropriation critique et responsable, conforme aux valeurs fondamentales de l'enseignement supérieur québécois.

## 3.3 Le métier étudiant : développer l'autonomie et la responsabilité face à l'IAg

Le cadre conceptuel d'*InitIAtion* repose également sur une reconnaissance explicite de la notion de métier d'étudiant, entendue non comme une simple condition administrative, mais comme un processus d'apprentissage à part entière. L'adaptation aux exigences de l'enseignement collégial requiert en effet l'acquisition progressive de compétences transversales, souvent peu formalisées dans les parcours éducatifs, mais essentielles à la réussite : planification du travail, gestion du temps, recherche et validation de sources, explicitation des raisonnements, autonomie et rigueur dans la prise de décision (CAPRES, 2020).

Dans ce contexte, la facilité d'accès aux outils d'IAg — dont les plus répandus sont dépourvus de toute visée pédagogique (Oudeyer, 2024) — modifie en profondeur les modalités de production et de transmission des savoirs (Roy *et al.*, 2025). Plus encore, ces technologies transforment les conditions mêmes dans lesquelles se développent les compétences associées au métier d'étudiant. En automatisant certaines tâches cognitives centrales, telles que la prise de notes, la rédaction de plans, la recherche documentaire ou l'explicitation d'une démarche, l'IAg peut fragiliser les processus d'apprentissage en court-circuitant les efforts cognitifs nécessaires à l'appropriation durable des savoirs (Abdelghani *et al.*, 2023). Ce risque de déresponsabilisation, inhérent à l'optimisation des tâches, s'accompagne toutefois d'une mise en tension des limites de l'automatisation, les étudiantes et étudiants étant confrontés à la nécessité de valider, nuancer, adapter, voire contester des réponses qui peuvent se révéler erronées ou fictives (Jabagi & Croteau, 2025).

La trousse *InitIAtion* adopte une approche nuancée : elle ne promeut ni l'usage généralisé de l'IAg ni son rejet. Elle vise à outiller les étudiantes et étudiants pour qu'ils identifient ce qui, dans leur apprentissage, relève d'un travail intellectuel irremplaçable. Le projet repose sur une posture lucide,

c'est-à-dire la capacité de comprendre les effets de l'IAg sur les apprentissages, la production des savoirs et l'autonomie propre au collégial. Cette lucidité critique les aide à comprendre les interactions entre les technologies qu'ils utilisent, les connaissances qu'ils construisent, et leur rôle comme sujets réflexifs, responsables et autonomes.

# 4 Modularité, adaptabilité et diffusion : les principes structurants d'*InitIAtion*

La trousse *InitIAtion* propose un parcours pédagogique progressif en trois modules d'environ 1h30 chacun, visant à développer une littératie critique et responsable de l'IAg. La progression suit la taxonomie de Bloom, des niveaux cognitifs de base (comprendre, appliquer) vers les plus complexes (analyser, évaluer, créer) [10].

Chaque module repose sur une logique de pérennité, fondée non sur des outils appelés à évoluer, mais sur des concepts fondamentaux de l'IA et de l'IAg (fonctionnement, entraînement, biais, limites, enjeux éthiques), ainsi que sur des compétences transversales essentielles au métier d'étudiant : organisation du travail, autonomie, jugement critique, évaluation de l'information, intégrité intellectuelle. Cet ancrage à la fois épistémologique et pédagogique garantit la durabilité, la transférabilité et la pertinence des apprentissages, quels que soient le champ de formation ou le programme d'études. Les contenus ont ainsi été pensés pour s'adapter à la diversité des parcours collégiaux, techniques ou préuniversitaires, en misant sur des compétences mobilisables dans différents contextes académiques et professionnels.

La scénarisation pédagogique repose sur une approche active, participative et réflexive. Chaque module est structuré autour de grands questionnements ouverts, auxquels les étudiantes et étudiants sont invités à répondre en mobilisant leurs connaissances, en expérimentant des outils et en échangeant en classe. Le parcours débute par l'exploration des représentations initiales et des usages spontanés, puis progresse vers l'analyse critique des contenus générés, la discussion collective et la prise de position argumentée. L'accent est mis sur l'apprentissage par l'action, dans une posture active et incarnée — les mains sur les touches — favorisant le développement du discernement, de l'autonomie et du jugement critique.

Les étudiantes et étudiants sont également encouragés à adapter ou concevoir leurs propres stratégies d'usage, ou à co-construire des repères en lien avec leur champ de formation. Cette dimension participative contribue au développement de leur agentivité numérique, entendue comme la capacité à agir de manière consciente, critique et contextualisée sur les outils et leur environnement (Ministère de l'Éducation, 2024).

La conception des modules repose sur une combinaison de ressources issues de documents pédagogiques existants (guides institutionnels, documents d'orientation), de recherches scientifiques récentes sur des enjeux spécifiques de l'IAg (hallucinations, biais cognitifs, anthropomorphisme par exemple), et d'outils conçus spécifiquement pour le contexte collégial. Cette approche a permis de bâtir un dispositif rigoureux, adapté aux exigences de l'enseignement supérieur, tout en répondant aux réalités concrètes du milieu (voir Annexe 1).

La trousse *InitIAtion* se déploie en trois modules complémentaires, articulés autour d'une progression

10. Taxonomie de Bloom révisée (domaine cognitif)

pédagogique cohérente (voir Annexe 1). Le premier module pose les fondements nécessaires à une compréhension critique de l'IAg. Il aborde le fonctionnement des grands modèles de langage (GML), leurs limites et la fiabilité des réponses produites, tout en questionnant l'idée, largement répandue chez les étudiantes et étudiants, selon laquelle l'IAg ne serait qu'un outil pour gagner du temps dans les tâches scolaires (KPMG, 2024). Les enjeux liés à la qualité de l'information, à l'intégrité intellectuelle et à la vie privée y sont introduits.

Le deuxième module approfondit l'analyse des usages, en outillant les étudiantes et étudiants à l'aide de la démarche CRISTAL. Cette posture, qui structure le module, les invite à être : Critiques, Responsables, Intègres, Sobres, Transparents, Autonomes, Libres et créatifs. À travers des discussions, des études de cas et des expérimentations, ils apprennent à interroger les réponses générées, évaluer les risques, assumer leurs choix et utiliser l'IAg de manière consciente, éthique et mesurée.

Le troisième module est consacré à la mise en pratique stratégique des apprentissages. Dans le cadre d'une tâche académique authentique, les personnes étudiantes mobilisent les principes de la posture CRISTAL pour formuler des requêtes efficaces, sélectionner les outils adaptés et utiliser l'IAg comme levier de réflexion et d'autonomie. L'agentivité numérique s'y consolide dans une approche critique, où l'IAg est intégrée comme soutien, et non comme substitut.

Les modules sont conçus pour une implantation flexible, selon les priorités pédagogiques, disciplinaires et les réalités propres à chaque programme. Ils peuvent s'intégrer à des cours disciplinaires, méthodologiques ou à des activités transversales. Cette modularité respecte les principes de l'enseignement supérieur québécois : liberté académique, autonomie professionnelle et institutionnelle (Conseil supérieur de l'éducation & Commission de l'éthique en science et en technologie, 2024).

*InitIAtion* n'est pas un cadre rigide, mais un fil conducteur structuré et évolutif. Chaque enseignante, enseignant ou programme peut adapter, enrichir ou transformer les modules selon ses objectifs, son contexte et le profil des étudiantes et étudiants. Cette logique ouverte respecte les valeurs du collégial : formation de citoyens engagés, diversité des parcours et accompagnement de la maturation intellectuelle (Conseil supérieur de l'éducation, 2019).

# 5 Recommandations pour une implantation réfléchie

Dès sa conception, *InitIAtion* a été pensée pour une diffusion large, progressive et interordres. Le projet vise à rejoindre l'ensemble des établissements collégiaux et universitaires du Québec, en priorisant les membres du PIM, qui soutient depuis 2018 des initiatives collaboratives en intelligence artificielle (Pôle interordres de Montréal, 2024). Sa diffusion s'appuiera sur la plateforme open source Moodle [11], où seront centralisés les contenus pédagogiques, les exercices et les guides à l'intention du personnel enseignant, facilitant ainsi la mutualisation des pratiques et leur adaptation locale.

Conformément aux principes de souplesse, de pérennité et de respect de l'autonomie institutionnelle, l'implantation d'*InitIAtion* ne repose pas sur un modèle unique. Sa modularité, son ancrage dans les compétences liées au métier d'étudiant, et sa compatibilité avec la diversité disciplinaire permettent une intégration contextualisée dans les programmes.

Il est recommandé d'introduire les modules aux moments clés du cycle de vie des programmes (élaboration, implantation, évaluation), périodes propices à une réflexion collective sur les contenus,

11. https://moodle.org/?lang=fr

les compétences transversales et l'actualisation pédagogique face aux transformations technologiques (Cégep de Saint-Laurent, 2024). Cette démarche s'inscrit dans l'approche programme, entendue comme une organisation cohérente et intégrée des apprentissages (Gosselin, 2021), et nécessite une concertation avec les comités de programme, garants de la qualité pédagogique et de l'adaptation aux besoins des étudiantes et étudiants.

Une formation précoce, dès l'entrée au collégial, est encouragée afin d'outiller rapidement les étudiantes et étudiants. En ce sens, une implantation graduelle sera amorcée à l'automne 2025 au Cégep de Saint-Laurent. Un protocole d'implantation est actuellement en co-construction avec la direction, et une démarche d'évaluation est en cours d'élaboration. Elle comprendra un questionnaire post-formation auprès des étudiantes et étudiants, ainsi que des entrevues avec le personnel enseignant. Cette implantation concertée reflète les valeurs fondatrices du projet *InitIAtion*, et ouvre une réflexion plus large sur ses retombées éducatives dans un environnement en constante évolution.

# 6 Conclusion

Le projet *InitIAtion* s'inscrit dans un contexte où les usages de l'IAg par les étudiantes et étudiants précèdent les cadres pédagogiques nécessaires pour les soutenir, les encadrer et les interroger. Face à cette asymétrie, la trousse propose une réponse réaliste, progressive et critique, non pour prescrire l'utilisation de ces outils, mais pour en accompagner les usages avec rigueur, discernement et responsabilité.

Ancrée dans les valeurs éducatives du réseau collégial québécois, *InitIAtion* repose sur une structure modulaire et ouverte, alignée sur les compétences transversales du métier d'étudier, les référentiels internationaux en littératie numérique, et les principes de liberté académique et d'autonomie professionnelle. Elle offre aux établissements des ressources adaptables à leurs besoins, contextes et priorités, tout en conservant une visée claire : développer l'agentivité, la pensée critique et l'éthique face aux technologies.

Ce projet, ambitieux, mais non prescriptif, tire sa pertinence de sa capacité à évoluer avec les milieux qui s'en emparent, à être adapté, questionné et enrichi. Il ne cherche pas à normaliser les usages de l'IAg, mais à fournir les moyens de les comprendre, les interroger et en évaluer les implications de façon éclairée et responsable. Dans un environnement numérique en constante mutation, *InitIAtion* vise à développer une prise de décision critique et autonome, fondée sur une analyse lucide des enjeux et des répercussions potentielles de l'IAg. C'est cette posture réflexive, durable et partagée qu'*InitIAtion* entend promouvoir, en cohérence avec les missions fondamentales de l'enseignement supérieur et les exigences d'une formation intellectuelle solide.

# Remerciements

Je tiens à exprimer ma profonde gratitude à mes collègues Michel Jean (département de philosophie) et Pier-Marc Gosselin (département d'informatique), avec qui cette réflexion sur l'intelligence artificielle en éducation a pris forme. C'est à travers nos échanges et nos premières expérimentations que l'idée d'*InitIAtion* a commencé à émerger, dans un esprit de collaboration interdisciplinaire que je m'efforce aujourd'hui de faire évoluer et de faire rayonner.

Je remercie sincèrement Alexandre P. Bédard, Julie Beaupré et Boris Nonveiller, coauteurs et coauteures de cette deuxième itération de la trousse *InitIAtion*. Leur rigueur, leur regard critique et leur volonté constante d'améliorer la qualité du projet ont permis de pousser cette initiative bien au-delà de ce que j'avais initialement imaginé. Merci au Carrefour d'innovation et de pédagogie universitaire de l'UQÀM ainsi qu'au Service des bibliothèques de l'UQÀM, qui ont rendu possible leur précieuse collaboration.

Je tiens également à remercier les enseignantes et enseignants du Cégep de Saint-Laurent qui ont participé à la phase exploratoire du projet, qui ont accepté de relever le défi mieux accompagner leur jeune public face aux enjeux liés à l'usage de l'IAg dans nos milieux éducatifs, et qui m'ont transmis des commentaires précieux. Un merci tout particulier aux étudiantes et étudiants, dont les questions, les essais, les hésitations et les élans ont constamment nourri la réflexion et orienté les ajustements réalisés.

Le projet *InitIAtion* n'aurait pu voir le jour sans le soutien du Cégep de Saint-Laurent, dont l'ouverture, la confiance et l'appui ont permis de concevoir, structurer et expérimenter la trousse dans un cadre pédagogique rigoureux. Je remercie également le Pôle interordres de Montréal (PIM) pour son soutien financier, ainsi que pour son rôle actif dans la création d'un espace de collaboration entre les ordres d'enseignement supérieur.

Merci à Daisy Le Corre pour la révision linguistique, qui a contribué à la clarté et à la cohérence de l'ensemble des documents, et à Alex Grenier, qui a su doter la trousse d'une identité visuelle forte, en phase avec les valeurs du projet.

Enfin, un merci chaleureux à toutes les personnes qui ont croisé le chemin de ce projet, au fil de discussions formelles ou informelles. Je pense entre autres à Philippe Soucy, analyste en informatique, et à Jean-Philippe Bourdon, bibliothécaire au cégep de Saint-Laurent, pour leurs conseils, leurs suggestions et leur accompagnement attentif.

# Références

ABDELGHANI R., SAUZÉON H. & OUDEYER P.-Y. (2023). Generative ai in the classroom : Can students remain active learners ?

ABRAN E. (2024). Sondage sur la population étudiante des cégeps (spec). Communication présentée dans le cadre de la rencontre REPTIC. 25 octobre. Disponible à l'adresse : https://reptic.ca/calendrier/rencontre-reptic-octobre-2024/.

ASSOCIATION QUÉBÉCOISE DE PÉDAGOGIE COLLÉGIALE (2023). Document de consultation sur l'utilisation des systèmes d'intelligence artificielle générative en enseignement supérieur : enjeux pédagogiques et éthiques. Association québécoise de pédagogie collégiale. Consulté à l'adresse : https://cdn.ca.yapla.com/company/CPYAEJBaj9LMVrlKzqeiCYmup/asset/files/MÃl'moire%20IA/AQPC_Memoire%20IA_2023.pdf.

CAPRES (2020). Le métier d'étudiant·e, de quoi parle-t-on ? Observatoire sur la réussite en enseignement supérieur. Consulté à l'adresse : https://oresquebec.ca/article-de-dossiers/notions-cles/le-metier-detudiante-de-quoi-parle-t-on-notion-cle/.

CENTRE CANADIEN POUR LA CYBERSÉCURITÉ (2023). L'intelligence artificielle générative. Gouvernement du Canada. Consulté à l'adresse : https://www.cyber.gc.ca/fr/

orientation/lintelligence-artificielle-generative-itsap00041.

CHAN C. K. Y. & HU W. (2023). Students' voices on generative ai : perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, **20**(1), 43. DOI : 10.1186/s41239-023-00411-8.

CHAN C. K. Y. & LEE K. K. W. (2023). The ai generation gap : Are gen z students more interested in adopting generative ai such as chatgpt in teaching and learning than their gen x and millennial generation teachers ? *Smart Learning Environments*, **10**(1), 60. DOI : 10.1186/s40561-023-00269-3.

COMITÉ PATRONAL DE NÉGOCIATION DES COLLÈGES (CPNC) (2024). Fneeq-csn convention collective 2023–2028. Consulté à l'adresse : https://fneeq.qc.ca/wp-content/uploads/CC_FNEEQ_2023-2028.pdf.

CONSEIL SUPÉRIEUR DE L'ÉDUCATION (2019). Les collèges après 50 ans : regard historique et perspectives. Gouvernement du Québec. Consulté à l'adresse : https://www.cse.gouv.qc.ca/wp-content/uploads/2019/05/50-0510-AV-colleges-apres-50-ans.pdf.

CONSEIL SUPÉRIEUR DE L'ÉDUCATION & COMMISSION DE L'ÉTHIQUE EN SCIENCE ET EN TECHNOLOGIE (2024). Intelligence artificielle générative en enseignement supérieur : enjeux pédagogiques et éthiques. Le Conseil ; La Commission. Consulté à l'adresse : https://www.cse.gouv.qc.ca/wp-content/uploads/2024/04/50-0566-RP-IA-generative-enseignement-superieur-enjeux-ethiques.pdf.

CÉGEP DE SAINT-LAURENT (2022a). 360-reu-01 réussir au collégial. Cégep de Saint-Laurent. 26 avril.

CÉGEP DE SAINT-LAURENT (2022b). Nouveau cours complémentaire au sujet de l'intelligence artificielle. Cégep de Saint-Laurent (Nouvelles). 25 mars. Consulté à l'adresse : https://www.cegepsl.qc.ca/nouvelle/nouveau-cours-complementaire-au-sujet-de-lintelligence-artificielle,

CÉGEP DE SAINT-LAURENT (2024). Politique institutionnelle de gestion et d'évaluation des programmes d'études (pigep). 12 juin. Consulté à l'adresse : https://www.cegepsl.qc.ca/documents/pigep.pdf.

DUFOUR I. (2023). Bilan du cours *Réussir au collégial* automne 2022-hiver 2023. Communication présentée au Rencontre du comité de la réussite. Communication orale.

DUFOUR I. & TARDIF I. (2023). Le cours *Réussir au collégial* : une mesure de soutien pour la transition secondaire-collégial. Communication présentée au Colloque de l'AQPC. Communication orale.

FENGCHUN M. & SHIOHIRA K. (2024). Ai competency framework for students. UNESCO. Consulté à l'adresse : https://doi.org/10.54675/JKJB9835, DOI : 10.54675/JKJB9835.

FENGCHUN M. & WAYNE H. (2024). Orientations pour l'intelligence artificielle générative dans l'éducation et la recherche. UNESCO. Consulté à l'adresse : https://doi.org/10.54675/HBCX3851, DOI : 10.54675/HBCX3851.

GOSSELIN P.-M., JEAN M. & JOUSSEMET F. (2024). Recommandations à la direction des Études et à la direction des ressources technologiques. Document interne, Cégep de Saint-Laurent. Document non publié.

GOSSELIN S. (2021). Formation ordinaire : Portrait de la formation collégiale. Conseil supérieur de l'éducation. Consulté à l'adresse : https://www.cse.gouv.qc.ca/wp-content/uploads/2021/02/50-2115-ER-Formation-collegiale-portrait.pdf.

GROUPE DE TRAVAIL SUR LA SENSIBILISATION DU PUBLIC (2022). Apprendre ensemble pour une intelligence artificielle responsable. Innovation, Sciences et Dévelop-

pement économique Canada. Consulté à l'adresse : https://ised-isde.canada.ca/site/advisory-council-artificial-intelligence/sites/default/files/attachments/2023/apprendre_ensemble_pour_une_intelligence_artificielle_responsable_minisi_approved.pdf.

JABAGI N. & CROTEAU A.-M. (2025). L'intelligence artificielle (ia) : amie ou ennemie de la motivation des étudiants et étudiantes universitaires ? *Revue internationale des technologies en pédagogie universitaire*, **22**(1), 4. DOI : 10.18162/ritpu-2025-v22n1-04.

JOUSSEMET F. (2025). Bilan d'usage de la trousse pilote initiation a24-h25. Cégep de Saint-Laurent. Document interne.

JOUSSEMET F., DUPLESSIS V., JEAN M. & GOSSELIN P.-M. (2024a). Outils d'intelligence artificielle générative intéressants pour la recherche. Cégep de Saint-Laurent. Document interne.

JOUSSEMET F., JEAN M. & GOSSELIN P.-M. (2023). Utilisation et intérêt pour les modèles de langages (ex : Chatgpt) en contexte scolaire. résultats de la consultation des enseignant.es du cégep de saint-laurent. Cégep de Saint-Laurent. Document interne.

JOUSSEMET F., JEAN M. & GOSSELIN P.-M. (2024b). Trousse de formation initiation (version préliminaire). Cégep de Saint-Laurent. Document interne.

JOUSSEMET F., JEAN M. & GOSSELIN P.-M. (2024c). Utilisation et intérêt pour les modèles de langages (ex : Chatgpt) en contexte scolaire. résultats de la consultation des étudiants.es du cégep de saint-laurent. Cégep de Saint-Laurent. Document interne.

JOUSSEMET F. & MEURS M.-J. (2024). L'ia générative en enseignement : des pistes de réflexion. Conférence d'ouverture présentée au colloque « Les opportunités pédagogiques de l'IA générative en enseignement supérieur : mirages et réalités ». [Conférence d'ouverture].

KPMG (2024). Les étudiants qui utilisent l'ia générative avouent qu'ils n'apprennent pas autant. KPMG. Consulté le 2 juin 2025.

LAFLEUR T. (2019). Le spec, un levier pour la réussite. Portail du réseau collégial du Québec. Consulté le 2 juin 2025.

MINISTÈRE DE L'ÉDUCATION (2024). L'utilisation pédagogique, éthique et légale de l'intelligence artificielle générative — guide destiné au personnel enseignant — 2024-2025. Gouvernement du Québec. Document officiel.

MINISTÈRE DE L'ÉDUCATION ET DE L'ENSEIGNEMENT SUPÉRIEUR (2018). Activités de mise à niveau et activités favorisant la réussite — Établissements d'enseignement collégial francophones. Gouvernement du Québec. Rapport officiel.

ORES (2024). Intelligence artificielle générative : qu'en disent les étudiantes et étudiants ? Observatoire sur la réussite en enseignement supérieur. Consulté le 2 juin 2025.

OUDEYER P.-Y. (2024). Ia générative, société et éducation : En quoi l'ia générative représente-elle un enjeu dans la formation des citoyens ? Conférence de consensus « Nouveaux savoirs et nouvelles compétences des jeunes » du Cnesco. Consulté le 2 juin 2025.

PETERS M. (2023). Note éditoriale : Intelligence artificielle et intégrité académique peuvent-elles faire bon ménage ? *Revue des sciences de l'éducation*, **49**(1), 1107846ar. DOI : 10.7202/1107846ar.

PÔLE INTERORDRES DE MONTRÉAL (2024). Invitation — manifestation d'intérêt. activités et outils pour l'utilisation judicieuse de l'ia générative en enseignement supérieur. Consulté le 2 juin 2025.

ROY N., PROUST-ANDROWKHA S., GRUSLIN E., VALLERAND V. & CHARLES E. (2025). L'intelligence artificielle au postsecondaire : entre enthousiasme et méfiance — introduction au numéro thématique. *Revue internationale des technologies en pédagogie universitaire*, **22**(1), 1. DOI : 10.18162/ritpu-2025-v22n1-01.

RÉGIS M. (2024). Bilan du cours réussir au collégial automne 2023-hiver 2024. Communication présentée à la Rencontre de la communauté du cours 360-RÉU.

VERCHÈRE C. (2024). Impacts et enjeux éthiques et sociaux des ia (dont chat gpt) [conférence d'ouverture]. Conférence d'ouverture communication présentée à Journée pédagogique Apprivoiser les IA depuis un an… et la suite ? Conférence d'ouverture.

WILSON L. O. (2016). Anderson and krathwohl bloom's taxonomy revised. Quincy College. Consulté le 2 juin 2025.

# Annexe 1 : Objectifs d'apprentissage des trois modules InitIAtion et matériel de conception

## Module 1 — Comprendre les fondements de l'IA générative (1h30)

**Objectifs d'apprentissage**

À la fin de ce module, les personnes étudiantes seront en mesure de :

- Identifier des outils d'IA utilisés dans leur quotidien.
- Reconnaître les caractéristiques qui distinguent l'IAg des autres formes d'IA.
- Expliquer ce que sont les grands modèles de langage (GML) et le rôle des données dans leur entraînement.
- Comprendre le fonctionnement de base d'un agent conversationnel et la logique probabiliste qui sous-tend ses réponses.
- Nommer les types de données récoltées par les IAg et réfléchir aux enjeux de confidentialité liés à leur usage.
- Prendre conscience des questions d'intégrité intellectuelle soulevées par l'usage des agents conversationnels.
- Discuter des limites acceptables dans l'usage de l'IAg en contexte scolaire, en lien avec les attentes en matière d'intégrité intellectuelle.

**Matériel de conception du module 1 :**


- Artificial Analysis. (2024, 15 septembre). AI Chatbot Comparison. Language Models.
- Artificial Analysis. (s. d.). Price USD per 1M Tokens. Comparison of Models: Intelligence, Performance & Price Analysis.
- Burnett, G. D. (2025, 26 avril). Will the humanities survive artificial intelligence? *The New Yorker*.
- Cégep de Saint-Laurent. (2023, 25 janvier). Politique institutionnelle d'évaluation des apprentissages (PIÉA).
- Centre canadien pour la cybersécurité. (2023, juillet). L'intelligence artificielle générative. Gouvernement du Canada.
- Cherrayil, N. K. (2025, 20 février). Which AI chatbot shares most data with third parties?
- Commission Nationale de l'Informatique et des Libertés. (s. d.a). Apprentissage par renforcement et rétroaction humaine. CNIL.
- Commission Nationale de l'Informatique et des Libertés. (s. d.b). Entraînement (ou apprentissage). CNIL.
- Elements of AI. (s. d.). Comment définir l'IA?
- International Center for Academic Integrity. (2021). *Fundamental Values of Academic Integrity* (3rd ed.).
- Jakhar, D. et Kaur, I. (2020). Artificial intelligence, machine learning and deep learning: Definitions and differences. *Clinical and Experimental Dermatology, 45*(1), 131–132.
- Joussemet, F., Jean, M. et Gosselin, P.-M. (2022). IA au quotidien : matériel d'implantation d'un nouveau cours complémentaire [moodle].
- Joussemet, F., Jean, M. et Gosselin, P.-M. (2024). *Trousse de formation InitIAtion* (version préliminaire). Cégep de Saint-Laurent.
- Joussemet, F. et Meurs, M.-J. (2024, 24 avril). L'IA générative en enseignement : des pistes

de réflexion [conférence d'ouverture]. Communication présentée à *Les opportunités pédagogiques de l'IA générative en enseignement supérieur : mirages et réalités*.
— OBVIA. (2025, janvier). *Glossaire de l'OBVIA*.
— Office québécois de la langue française. (2025). Agent conversationnel. Gouvernement du Québec.
— Russel, S., Perset, K. et Grobelnik, M. (2023, 29 novembre). Updates to the OECD's definition of an AI system explained. *OECD.AI*.
— Postel-Vinay (dir.). (2025). Pour une IA responsable et éthique. *Les Annales des Mines*, (29), 187.
— Raspberry Pi Foundation. (2024). Grands modèles de langage (GML). *Experience AI*.
— Thienot, É. (2024). ChatGPT — Decryptage [sketchnote].
— Tremblay, A. (2025). WordFlow AI.
— Université de Genève. (2024). *Guide à l'intention de la communauté universitaire. Intelligence artificielle générative*.
— Université de Montréal. (2018). *La Déclaration de Montréal pour un développement responsable de l'intelligence artificielle*.
— Vangrunderbeeck (dir.), P. (2024, octobre). *Intégrer l'IA générative dans les stratégies pédagogiques*. UCLouvain.

## Module 2 — Développer une posture critique avec la démarche CRISTAL (3 h)

**Objectifs d'apprentissage**

À la fin de ce module, les personnes étudiantes seront en mesure de :

— Distinguer une information factuelle d'une fabulation dans les réponses générées par un agent conversationnel ;
— Évaluer la neutralité d'un contenu généré et repérer les biais, stéréotypes ou normes sociales implicites qui y sont véhiculés ;
— Nommer les risques d'un usage non déclaré ou excessif de l'IAg en contexte scolaire, et appliquer les principes d'intégrité intellectuelle associés à son utilisation ;
— Expliquer les impacts environnementaux associés aux outils d'IAg et discuter des pratiques favorisant une utilisation numérique sobre et responsable ;
— Reconnaître les enjeux de vie privée liés au partage de données avec un agent conversationnel, et adopter les précautions nécessaires pour protéger les informations sensibles ;
— Décrire les mécanismes d'imitation du langage humain par les agents conversationnels (effet d'anthropomorphisme) et maintenir une distance critique face à ces interactions ;
— Explorer les possibilités créatives offertes par l'IAg tout en restant libre dans sa pensée, capable de juger, de sélectionner et de transformer les contenus proposés ;
— Mettre en pratique les principes de la démarche CRISTAL pour adopter un usage éthique, critique, autonome et réfléchi de l'IAg.

**Matériel de conception du module 2 :**

— Abdelghani, R., Sauzéon, H. et Oudeyer, P.-Y. (2023, 10 novembre). *Generative AI in the Classroom: Can Students Remain Active Learners?* arXiv.
— Audran, J. (2024). *Cinq enjeux d'évaluation face à l'émergence des IA génératives en éducation*. Mesure et évaluation en éducation, 47(1), 6–26.

— BBC Media Centre. (2025). *Representation of BBC News content in AI Assistants*.
— Borji, A. (2023, 3 avril). *A Categorical Archive of ChatGPT Failures*. arXiv.
— Bureau du droit d'auteur. (2025). *Intelligence artificielle*. Université Laval — Bibliothèque.
— Cégep de Saint-Laurent. (2023, 25 janvier). *Politique institutionnelle d'évaluation des apprentissages (PIÉA)*.
— Chan, C. K. Y. et Hu, W. (2023). *Students' voices on generative AI: perceptions, benefits, and challenges in higher education*. International Journal of Educational Technology in Higher Education, 20(1), 43.
— Chen, S. (2025). *How much energy will AI really consume? The good, the bad and the unknown*. Nature, 639(8053), 22–24.
— Commission de l'éthique en science et en technologie. (2021, 26 février). *L'effet rebond : la face cachée du bilan environnemental des technologies numériques*. Gouvernement du Québec.
— Commission de l'éthique en science et en technologie. (2025, 30 janvier). *IA générative : à quels coûts pour la planète?* Gouvernement du Québec.
— Conseil supérieur de l'éducation et Commission de l'éthique en science et en technologie. (2024). *Intelligence artificielle générative en enseignement supérieur : enjeux pédagogiques et éthiques*.
— Creely, E. et Blannin, J. (2025). *Creative partnerships with generative AI. Possibilities for education and beyond*. Thinking Skills and Creativity, 56, 101727.
— Curiale, T., Acquatella, F., Gros, L., Cosquer, M. et Tisseron, S. (2022). *L'anthropomorphisme, enjeu de performance pour les chatbots*. Revue internationale de psychosociologie et de gestion des comportements organisationnels, 28(72), 101–123.
— Deng, R., Jiang, M., Yu, X., Lu, Y. et Liu, S. (2025). *Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies*. Computers & Education, 227, 105224.
— Doshi, A. R. et Hauser, O. P. (2024). *Generative AI enhances individual creativity but reduces the collective diversity of novel content*. Science Advances, 10(28), eadn5290.
— Fan, Y. et al. (2024). *Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance*. British Journal of Educational Technology.
— Fengchun, M. et Wayne, H. (2024). *Orientations pour l'intelligence artificielle générative dans l'éducation et la recherche*. UNESCO.
— Fournier-Tombs, E. et Castets-Renard, C. (2024). *Algorithmes et propagation de normes culturelles sexospécifiques*. Dans C. Brin & V. Guèvremont (dir.), *Intelligence artificielle, culture et médias* (pp. 429–448). PUL.
— Hwang, Y. et Wu, Y. (2025). *The influence of generative artificial intelligence on creative cognition of design students*. Frontiers in Psychology, 15, 1455015.
— International Center for Academic Integrity. (2021). *Fundamental Values of Academic Integrity* (3rd ed.).
— Jabagi, N. et Croteau, A.-M. (2025). *L'intelligence artificielle (IA) : amie ou ennemie de la motivation des étudiants et étudiantes universitaires ?* Revue internationale des technologies en pédagogie universitaire, 22(1), 4.
— Jaźwińska, K. et Chandrasekar, A. (2024, 27 novembre). *How ChatGPT Search (Mis)represents Publisher Content*. Columbia Journalism Review.
— KPMG. (2024, 21 octobre). *Les étudiants qui utilisent l'IA générative avouent qu'ils n'apprennent pas autant*. KPMG.

— Ministère de l'Éducation et de l'Enseignement supérieur. (2017). *Composantes de la formation générale*. Gouvernement du Québec.
— Office québécois de la langue française. (2004). *Effet de halo*. Gouvernement du Québec.
— Oudeyer, P.-Y. (2024, octobre). *IA générative, société et éducation : En quoi l'IA générative représente-elle un enjeu dans la formation des citoyens?* Conférence du Cnesco.
— Perrin, N., Piot, D., Vita, L., Bationo-Tillon, A. et Guibourdenche, J. (2025). *Des hypothèses pour concevoir des tâches permettant aux étudiants et étudiantes d'évaluer la pertinence des textes générés par les IA*. Revue internationale des technologies en pédagogie universitaire, 22(1), 6.
— Phare. (2025, 14 avril). *Phare LLM Benchmark*. Phare.
— Seeger, A.-M., Pfeiffer, J. et Heinzl, A. (2021). *Texting with Humanlike Conversational Agents: Designing for Anthropomorphism*. Journal of the Association for Information Systems, 22(4), 931–967.
— UQÀM. (s.d.). *ChatGPT et intelligence artificielle générative*. Service des bibliothèques.
— Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J. et Fedus, W. (2024, 7 novembre). *Measuring short-form factuality in large language models*. arXiv.
— You, J. (2025, 7 février). *How much energy does ChatGPT use?* Epoch AI.
— Yusuf, A., Pervin, N. et Román-González, M. (2024). *Generative AI and the future of higher education: a threat to academic integrity or reformation?* International Journal of Educational Technology in Higher Education, 21(1), 21.

## Module 3 — Formuler des requêtes et utiliser l'IAg avec discernement (1h30)

**Objectifs d'apprentissage**

À la fin de ce module, les personnes étudiantes seront en mesure de :

— Appliquer des stratégies de rédactique alignées sur sa posture d'apprentissage, en s'appuyant sur les repères de la méthode CRISTAL pour formuler des requêtes ciblées, claires et efficaces ;
— Utiliser un outil d'IAg basé sur la génération augmentée de récupération (RAG) en tenant compte de ses fonctionnalités ;
— Élaborer une stratégie d'usage des agents conversationnels en fonction de ses besoins réels et des conditions d'usage ;
— Créer un outil ou une ressource personnalisée à l'aide d'un agent conversationnel pour soutenir son apprentissage de façon autonome, critique et responsable.

**Matériel de conception du module 3 :**

— Agentic AI : A Progression of Language Model Usage [youtube]. (2025, 24 janvier).
— Délégation Régionale Académique au Numérique Éducatif (DRANE). (2024, 7 novembre). *Quel est l'impact environnemental d'une IA générative?* Région Académique Île-de-France.
— Google. (2025a). *Découvrez comment NotebookLM protège vos données.* Aide NotebookLM.
— Google. (2025b). *Premiers pas avec NotebookLM et NotebookLM Plus.* Aide NotebookLM.
— Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. et Iwasawa, Y. (2023, 29 janvier). *Large Language Models are Zero-Shot Reasoners.* arXiv.
— Le Collimateur. (2024, 23 janvier). *La « génération augmentée de récupération », vous connaissez?*

— Lepage, A. et Roy, N. (2025). *Le développement d'une échelle de mesure de la littératie de l'intelligence artificielle chez les enseignants et les enseignantes du postsecondaire.* Mesure et évaluation en éducation, 47(2), 39–69.
— Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., et al. (2021, 12 avril). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.* arXiv.
— Lopez, C. (2024, 26 octobre). *Comment utiliser l'IA de manière plus douce pour le climat.* Le Devoir.
— Madaan, A., Tandon, N., Gupta, P., Hallinan, S., et al. (2023, 25 mai). *Self-Refine : Iterative Refinement with Self-Feedback.* arXiv.
— Ministère de la Culture. (s. d.). *Compar:ia. Le comparateur d'IA conversationnelles.*
— OpenAI. (s. d.a). *Reasoning best practices.* OpenAI Platform.
— OpenAI. (s. d.b). *Reasoning models.* OpenAI Platform.
— OpenAI. (s. d.c). *Text generation and prompting.* OpenAI Platform.
— Raiza, M. et Johnson, S. (2023, 12 juillet). *Introducing NotebookLM.*
— Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., et al. (2024, 30 décembre). *The Prompt Report: A Systematic Survey of Prompting Techniques.* arXiv.
— Si, C., Gan, Z., Yang, Z., Wang, S., et al. (2023, 15 février). *Prompting GPT-3 To Be Reliable.* arXiv.
— Université de Genève. (2024). *Guide à l'intention de la communauté universitaire. Intelligence artificielle générative.*
— Université de Sherbrooke. (2025, 6 mai). *Intelligences artificielles : Génération de texte enrichie d'information.* Bibliothèques et archives.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Intégration encadrée de l'IA générative dans une activité d'apprentissage par problème en école d'ingénieur

Christophe Tilmant[1] Susan Arbon-Leahy[2]
(1) Université Clermont Auvergne, Clermont Auvergne INP, CNRS, Institut Pascal, F-63000 Clermont–Ferrand, France
(2) Université Clermont Auvergne, Clermont Auvergne INP, ISIMA, F-63000 Clermont–Ferrand, France
christophe.tilmant@uca.fr, susan.arbon-leahy@uca.fr

## RÉSUMÉ

Cet article présente une expérimentation pédagogique menée dans une école d'ingénieurs en informatique visant à encadrer l'usage de l'Intelligence Artificielle Générative (IAG) dans le cadre d'une activité d'apprentissage par problème. Intégrée à une Situation d'Apprentissage et d'Évaluation (SAÉ) de première année (niveau L3), l'activité proposée s'appuie sur l'analyse d'un brevet en anglais décrivant un algorithme de reconnaissance musicale de type Shazam. À partir de ce document, les étudiants sont amenés à produire un glossaire technique bilingue, en interaction réflexive avec une IAG. Le dispositif vise à développer conjointement des compétences disciplinaires (traitement du signal, mathématiques appliquées), linguistiques (anglais scientifique) et transversales (analyse distanciée, réflexivité sur l'usage des outils d'IAG). Les résultats observés montrent une forte implication des étudiants, une qualité linguistique et technique des productions, ainsi qu'une capacité à identifier et discuter les apports et limites de l'IAG. Cette activité constitue une première étape vers une intégration systémique et réfléchie des IAG dans les cursus d'ingénierie.

## ABSTRACT

**Structured Integration of Generative AI in a Problem-Based Learning Activity in Engineering Education**

This paper presents a pedagogical experiment conducted in a computer engineering school to frame the use of generative artificial intelligence (GAI) in a problem-based learning activity. Embedded in a first-year project, students were asked to analyze a scientific patent written in English describing a music recognition algorithm (Shazam). Based on this reading, they produced a bilingual technical glossary, using GAI as a tool for feedback and reflection. The task was designed to foster disciplinary knowledge (signal processing, applied mathematics), linguistic skills (technical English), and critical thinking. Results indicate a strong engagement, improved quality of scientific writing, and the ability to evaluate GAI outputs. This activity serves as a foundation for future curricular integration of GAI in engineering education.

MOTS-CLÉS : IA générative, apprentissage par problème, traitement du signal, anglais scientifique, pédagogie critique, transformation curriculaire..

KEYWORDS: generative AI, problem-based learning, signal processing, scientific English, critical thinking, curriculum innovation..

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 63–69.

# 1 Introduction

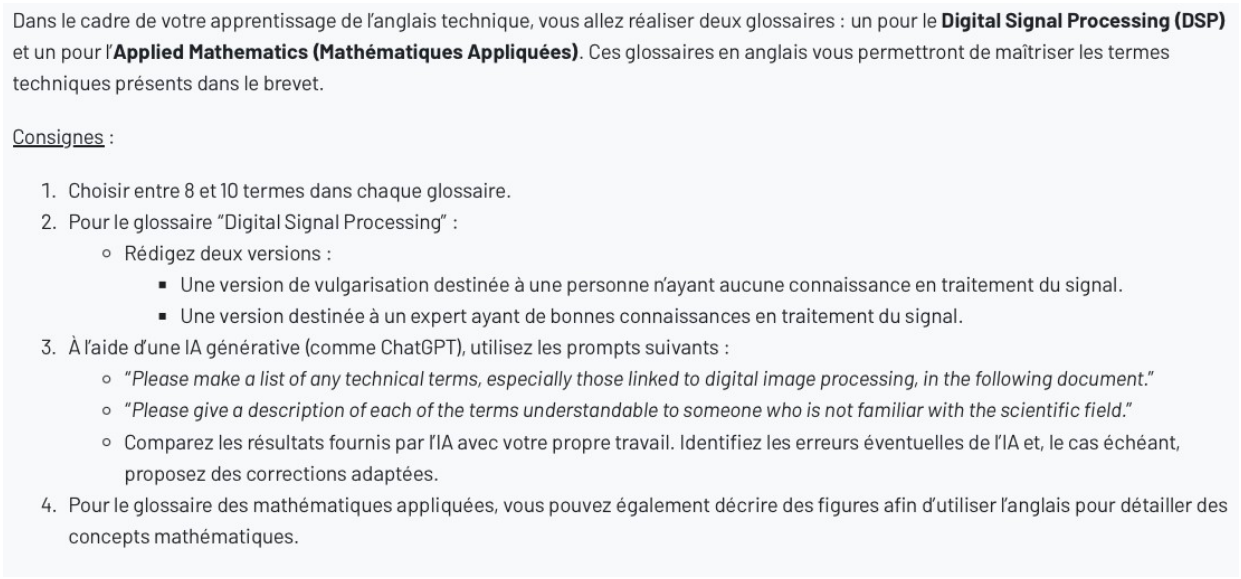
L'essor rapide des intelligences artificielles génératives (IAG) dans l'enseignement supérieur soulève de nouveaux enjeux pédagogiques majeurs. Comment former les étudiants à utiliser ces outils de manière éclairée et critique ? Comment éviter que leur usage ne vienne court-circuiter les apprentissages, en substituant à l'effort cognitif des réponses générées, souvent opaques ou approximatives ? Dans le cadre d'une formation d'ingénieur fondée sur une Approche Par Compétences (APC), ces interrogations prennent une dimension particulière. Les activités pédagogiques y reposent sur des situations authentiques, appelées Situations d'Apprentissage et d'Évaluation (SAÉ) (Tardif, 2006), qui placent les étudiants en position d'acteurs autonomes de leur apprentissage. Or, cette autonomie, sans cadre méthodologique explicite, peut conduire à un usage mécanique ou non critique des IAG, au détriment des objectifs d'apprentissage visés (Cao & Dede, 2023). Par « critique », nous entendons ici une capacité à interroger les outils numériques de manière argumentée, distanciée et contextualisée. Face à ces risques, nous avons conçu une activité pédagogique intégrée à une SAÉ disciplinaire, dans laquelle l'usage de l'IAG est structuré, encadré et réfléchi. L'ambition est claire : faire de l'IAG non une béquille automatisée, **mais un vecteur de réflexion, de clarification conceptuelle et de développement de la pensée critique**. Cette expérimentation vise ainsi à initier, dès la première année, une posture de dialogue raisonné avec les IAG, ancrée dans les enjeux cognitifs, techniques et linguistiques propres à une formation scientifique exigeante. Cet article rend compte d'un **retour d'expérience structuré**, visant à documenter la mise en œuvre d'un dispositif pédagogique mobilisant l'IAG, sans prétendre à une évaluation expérimentale contrôlée.

# 2 Contexte : une SAÉ centrée sur la reconnaissance musicale

L'activité expérimentale s'inscrit dans une SAÉ destinée à des étudiants de première année d'école d'ingénieur (niveau L3), au sein d'une formation en informatique. Cette SAÉ a pour objectif la modélisation d'un système de reconnaissance musicale automatique, fondé sur le principe du calcul d'empreintes acoustiques invariantes à des transformations sonores, et de leur recherche efficace dans une base de données. Le dispositif est directement inspiré du brevet de l'algorithme Shazam (Wang, 2003), utilisé comme support central de l'activité. L'un des principaux obstacles rencontrés par les étudiants est la lecture de ce brevet, rédigé en anglais scientifique dense et technique. Celui-ci mobilise un ensemble de concepts mathématiques et algorithmiques avancés : transformée de Fourier, spectrogrammes, corrélation d'empreintes, robustesse au bruit, etc. Cette lecture constitue un défi pédagogique significatif, sollicitant de manière transversale des compétences en mathématiques appliquées, en algorithmique et en anglais scientifique. Par son exigence conceptuelle et linguistique, cette tâche active plusieurs compétences au cœur de la formation : compréhension fine de documents techniques en langue étrangère, la mobilisation de modèles mathématiques pour le traitement du signal, ainsi que la capacité à formaliser une architecture logicielle à partir d'un document existant. Dans cette perspective, le dispositif expérimental s'inscrit dans une dynamique de montée en complexité cognitive, que l'on peut analyser au travers de la **taxonomie révisée de Bloom** (Anderson & Krathwohl, 2001). Les tâches proposées mobilisent plusieurs niveaux cognitifs : compréhension des concepts (niveau *Understand*) et leur application par la reformulation technique (niveau *Apply*).

C'est dans ce cadre que l'usage d'une IAG a été introduit, non pas comme un substitut à la compréhension, mais comme **outil d'accompagnement raisonné**. L'IAG a été pensée ici comme un **médiateur cognitif**, capable de soutenir l'explicitation des concepts clés (notamment autour des

notions de Fourier, d'invariance et de spectrogrammes), tout en facilitant une première analyse linguistique du document source. Cependant, cette activité pédagogique doit être revisitée à la lumière des propositions récentes visant à actualiser la **taxonomie de Bloom à l'ère de l'IA** [1]. Ces travaux intègrent désormais des compétences supérieures telles que la régulation de la qualité des résultats produits par l'IA, ainsi que la conscience éthique liée à l'usage algorithmique. Le glossaire argumenté, proposé aux étudiants, les engage ainsi dans **un parcours d'apprentissage significatif** (*meaningful learning*), soutenu par l'IAG sans s'y substituer. Cette démarche favorise le développement d'une autonomie cognitive réflexive et informée.



Dans le cadre de votre apprentissage de l'anglais technique, vous allez réaliser deux glossaires : un pour le **Digital Signal Processing (DSP)** et un pour l'**Applied Mathematics (Mathématiques Appliquées)**. Ces glossaires en anglais vous permettront de maîtriser les termes techniques présents dans le brevet.

Consignes :

1. Choisir entre 8 et 10 termes dans chaque glossaire.
2. Pour le glossaire "Digital Signal Processing" :
    - Rédigez deux versions :
        - Une version de vulgarisation destinée à une personne n'ayant aucune connaissance en traitement du signal.
        - Une version destinée à un expert ayant de bonnes connaissances en traitement du signal.
3. À l'aide d'une IA générative (comme ChatGPT), utilisez les prompts suivants :
    - *"Please make a list of any technical terms, especially those linked to digital image processing, in the following document."*
    - *"Please give a description of each of the terms understandable to someone who is not familiar with the scientific field."*
    - Comparez les résultats fournis par l'IA avec votre propre travail. Identifiez les erreurs éventuelles de l'IA et, le cas échéant, proposez des corrections adaptées.
4. Pour le glossaire des mathématiques appliquées, vous pouvez également décrire des figures afin d'utiliser l'anglais pour détailler des concepts mathématiques.

FIGURE 1 – Consignes pédagogiques encadrant la production du glossaire argumenté

# 3 Dispositif pédagogique : glossaire technique bilingue et interaction avec l'IAG

Le livrable attendu (cf. Figure 1) prend la forme d'un **glossaire argumenté** portant sur 8 à 10 termes extraits du brevet Shazam, choisis pour leur pertinence dans les domaines du traitement du signal ou des mathématiques appliquées. L'activité est conçue selon une progression en trois étapes, visant à articuler production personnelle, interaction avec une IAG, et analyse distanciée.

1. **Définition initiale** : Les étudiants rédigent pour chaque terme deux définitions distinctes : une version vulgarisée (destinée à un public non spécialiste) et une version experte (destinée à un lecteur scientifique). Ces définitions s'appuient sur leurs acquis disciplinaires (mathématiques, traitement du signal) ainsi que sur les cours d'anglais technique.
2. **Interaction avec l'IAG** : Chaque paire de définitions est ensuite soumise à une IAG à l'aide d'un prompt prédéfini (Marvin *et al.*, 2024), conçu pour solliciter une correction, amélioration et justification de la part du modèle (Walter, 2024). L'IAG propose alors une reformulation accompagnée d'explications sur les modifications suggérées.

1. https://ecampus.oregonstate.edu/faculty/artificial-intelligence-tools/blooms-taxonomy-revisited/

3. **Analyse distanciée** : Les étudiants sont invités à comparer leurs définitions avec celles générées par l'IAG, et à en tirer une **analyse réflexive**. Quelles différences observe-t-on entre les deux versions ? Quels éléments apportés par l'IAG sont utiles ou, au contraire, discutables ? L'IAG a-t-elle commis des erreurs ou introduit des ambiguïtés ?

Cette étape engage les étudiants dans un véritable exercice de **métacognition**, les amenant à porter un regard distancié sur leurs productions et sur la pertinence des propositions générées. Ce dispositif suit une logique **tripartite** : **production** → **interaction** → **évaluation critique**, favorisant une posture active et réflexive vis-à-vis de l'outil. Il mobilise un large spectre de compétences : rédaction scientifique, maîtrise des concepts disciplinaires, expression en anglais technique et jugement réflexif face à une technologie algorithmique. L'ensemble s'inscrit dans une démarche d'apprentissage **encadrée**, **critique** et **contextualisée** de l'usage des IAG en contexte académique exigeant, en cohérence avec les recommandations récentes sur l'intégration pédagogique de ces technologies (Walter, 2024).

# 4 Résultats qualitatifs observés : apprentissages disciplinaires, linguistiques et réflexifs

Les productions des étudiants ont été évaluées de manière qualitative selon une méthode croisée, par deux enseignants (l'un en mathématiques appliquées, l'autre en anglais scientifique), à partir des **attendus pédagogiques définis dans la SAÉ** et donc du **référentiel de compétences** associé. Cette évaluation a porté sur la qualité technique, linguistique et réflexive des productions. Plusieurs constats significatifs se dégagent de cette analyse croisée, illustrant les effets pédagogiques du dispositif.

— **Appropriation conceptuelle** : les définitions rédigées par les étudiants révèlent une bonne compréhension des notions fondamentales, telles que spectrogram, fingerprint ou cepstral coefficient. La confrontation avec les propositions de l'IAG permet souvent d'enrichir ou de préciser ces notions, tout en identifiant d'éventuelles erreurs.
— **Progression linguistique** : les définitions en anglais gagnent en précision, en fluidité syntaxique et en cohérence lexicale, en particulier après l'étape d'interaction avec l'IAG. Les étudiants améliorent leur maîtrise des tournures spécifiques au discours scientifique.
— **Esprit critique** : les analyses produites montrent une capacité à évaluer la pertinence des réponses de l'IAG, à en identifier les limites, et à proposer des reformulations mieux adaptées au contexte.

Exemples illustratifs issus des productions étudiantes :

— **Exemple 1** - Terme "*spectrogram*" : Dans leur version initiale, les étudiants décrivent un spectrogramme comme une représentation graphique du contenu fréquentiel d'un signal audio au cours du temps. L'IAG propose une version enrichie incluant les notions de "*sliding window*" et de "*frequency bins*". Les étudiants reconnaissent la pertinence de ces ajouts, tout en soulignant leur complexité pour une définition vulgarisée. Ils proposent en réponse un exemple imagé pour renforcer la compréhension, montrant ainsi leur capacité à adapter la définition au public visé.
— **Exemple 2** - Terme "*cepstral coefficient*" : L'IAG confond ici les coefficients cepstraux avec les MFCC (*Mel Frequency Cepstral Coefficients*), utilisés principalement en reconnaissance vocale. Les étudiants identifient cette erreur conceptuelle et reformulent leur définition en

précisant que, dans le contexte musical, le cepstre résulte de l'application de la transformée de Fourier inverse sur le logarithme du spectre. Ils soulignent que cette opération permet d'identifier des motifs périodiques correspondant aux harmoniques.

— **Exemple 3** -Travail sur le terme "*pitch-corrected tempo variation*" : L'IAG a proposé une définition correcte mais trop généraliste. L'étudiant a reformulé : "*It refers to a technique allowing to change the speed of playback without altering the pitch, by adjusting the time domain while preserving frequency content*". L'enseignante d'anglais a salué l'usage approprié de la terminologie ("*playback*", "*frequency content*").

— **Exemple 4** - Terme "*fingerprint*" : La définition initiale produite est : "*A fingerprint is a condensed representation of the audio signal which preserves enough unique traits to allow identification.*" L'IAG y ajoute les notions de "*hashing techniques*" et de "*robustness to noise*". L'étudiant intègre ces éléments, tout en ajustant la formulation pour éviter la redondance. Il ajoute une clarification sur le terme "*hash*" : "*a mathematical compression function*", démontrant sa capacité à vulgariser un concept complexe sans le dénaturer.

— **Exemple 5** - Maîtrise de l'anglais scientifique : Plusieurs étudiants montrent des progrès notables dans l'utilisation de structures typiques du discours scientifique, telles que "*It is defined as...*", "*This concept allows to...*", ou "*In the context of signal processing...*". L'enseignante remarque également un meilleur usage des connecteurs logiques et des justifications de type "*because it ensures time-invariance*". Dans un cas, un étudiant corrige un emploi incorrect du mot "*mean*" (voulu au sens de "*average*"), non détecté par l'IAG, illustrant une vigilance linguistique autonome.

Ces exemples montrent que l'IAG, loin de court-circuiter l'apprentissage, joue ici un rôle de **catalyseur**. Elle permet aux étudiants d'exercer leur discernement, de formuler des choix argumentés, et de tester la robustesse d'un outil désormais omniprésent. Ce dispositif favorise ainsi un développement simultané des **compétences disciplinaires**, **langagières** et **réflexives**, dans une perspective d'usage distancié et éclairé des technologies d'IAG.

# 5 Discussion : originalité et transférabilité du dispositif

Les productions étudiantes ont été jugées de **très bonne qualité** par les deux enseignants évaluateurs. La structure tripartite du dispositif a **mis en lumière le raisonnement cognitif des apprenants** et leur capacité à identifier, commenter et corriger les erreurs de l'IAG. L'évaluation croisée a mis en évidence une double progression : sur le plan disciplinaire, une meilleure formalisation des concepts liés au traitement du signal ; sur le plan linguistique, une amélioration de la précision lexicale, de la structuration syntaxique et de l'adaptation du registre en fonction du destinataire. L'activité a également généré un **effet retour sur les pratiques pédagogiques** : l'enseignante d'anglais a pu repérer des erreurs récurrentes à partir des glossaires, qu'elle a réinvesties dans l'ajustement de son programme.

Le dispositif se distingue par son articulation explicite entre production humaine, intervention algorithmique et analyse distanciée ; son intégration fluide dans une SAÉ existante ; la production d'un livrable structuré et traçable ; et une initiation structurée à un usage raisonné, éthique et professionnel des IAG. Contrairement à des approches centrées sur la simple génération automatisée, l'IAG est ici interrogée, mise à l'épreuve et discutée. Elle devient un outil d'explicitation technique et un vecteur de clarification, conformément aux principes de la pédagogie critique du numérique (Walter, 2024) et

de l'AI Literacy (Ng *et al.*, 2021). Ce type d'activité illustre également une **relecture pédagogique des objectifs d'apprentissage**. La taxonomie de Bloom revisitée dans le contexte des outils d'IA[2] invite à considérer non seulement les niveaux cognitifs traditionnels, mais aussi la capacité à articuler savoirs et outils technologique. La capacité à interagir avec une IAG devient ainsi un marqueur de compétences.

Malgré ces résultats encourageants, nous reconnaissons que **cette première expérimentation n'a pas été instrumentée par des outils d'évaluation systématiques**, tels que des journaux de bord, des questionnaires ou des entretiens structurés. Cette limite méthodologique est assumée dans la mesure où l'objectif principal était l'exploration d'un usage encadré de l'IAG en situation authentique. Bien que cette activité repose sur une expérimentation locale, sa structure modulaire et son articulation claire entre production humaine, intervention algorithmique et analyse réflexive en font un **modèle transférable**. Elle pourrait être adaptée à d'autres domaines, notamment en formation technique, linguistique ou en sciences humaines. Il convient toutefois de signaler un **point de vigilance éthique et écologique** : si les étudiants utilisent aujourd'hui spontanément des outils d'IAG dans leurs travaux, y compris en dehors des temps d'encadrement pédagogique, l'impact environnemental de ces usages reste une question ouverte. Cet aspect, non traité dans la présente expérimentation, mérite d'être pris en compte dans une réflexion plus large. Ce type d'activité, à la fois formative, réflexive et facilement intégrable à des contextes existants, constitue une **entrée opérationnelle et évolutive** pour accompagner la transformation curriculaire de l'enseignement supérieur face aux enjeux éducatifs posés par l'essor des IAG.

# 6 Conclusion

L'activité présentée a permis d'atteindre plusieurs objectifs pédagogiques complémentaires. Les étudiants se sont approprié des notions complexes liées au traitement du signal, ont enrichi leur expression scientifique en anglais, et ont développé un regard distancié sur les réponses générées par une IAG. Les retours qualitatifs recueillis, tant du point de vue des enseignants que des productions rendues, confirment la pertinence de ce format structurant. Loin d'un usage passif ou délégatif de l'IAG, les étudiants sont ici engagés dans un véritable processus réflexif. Ils produisent, comparent, analysent, et justifient. La confrontation à l'IAG devient alors un levier pour penser plus précisément, formuler plus rigoureusement, et affirmer une posture d'auteur. Si des usages détournés restent théoriquement possibles, le contexte pédagogique dans lequel cette activité a été menée (promotion de taille restreinte, encadrement régulier, projet intégré) limite largement ce risque. Introduire dès la première année un usage critique et encadré des IAG constitue une première étape vers une intégration curriculaire raisonnée et progressive de ces outils. Cette expérimentation ouvre des perspectives concrètes pour la suite du parcours de formation. En deuxième année, les étudiants seront amenés à concevoir leurs propres prompts (Marvin *et al.*, 2024), favorisant une montée en autonomie dans la maîtrise des IG. À plus long terme, l'extension de ce type d'approche à d'autres SAÉ ou modules disciplinaires est envisagée, tout comme une réflexion plus large à l'échelle de la composante sur l'évolution des référentiels de formation face aux enjeux sociotechniques portés par l'intelligence artificielle.

2. https://ecampus.oregonstate.edu/faculty/artificial-intelligence-tools/blooms-taxonomy-revisited/

# Références


ANDERSON L. W. & KRATHWOHL D. R. (2001). *A taxonomy for learning, teaching, and assessing : A revision of Bloom's taxonomy of educational objectives : complete edition*. Addison Wesley Longman, Inc.

CAO L. & DEDE C. (2023). Navigating a world of generative ai : Suggestions for educators. *The next level lab at harvard graduate school of education*, **5**(2).

MARVIN G., HELLEN N., JJINGO D. & NAKATUMBA-NABENDE J. (2024). Prompt engineering in large language models. In I. J. JACOB, S. PIRAMUTHU & P. FALKOWSKI-GILSKI, Éds., *Data Intelligence and Cognitive Informatics*, p. 387–402, Singapore : Springer Nature Singapore. DOI : 10.1007/978-981-99-7962-2_30.

NG D. T. K., LEUNG J. K. L., CHU S. K. W. & QIAO M. S. (2021). Conceptualizing ai literacy : An exploratory review. *Computers and Education : Artificial Intelligence*, **2**, 100041. DOI : 10.1016/j.caeai.2021.100041.

TARDIF J. (2006). *L'évaluation des compétences : documenter le parcours de développement*. Chenelière éducation,.

WALTER Y. (2024). Embracing the future of artificial intelligence in the classroom : the relevance of ai literacy, prompt engineering, and critical thinking in modern education. *International Journal of Educational Technology in Higher Education*, **21**(1), 15. DOI : 10.1186/s41239-024-00448-3.

WANG A. (2003). An industrial strength audio search algorithm. In *Ismir*, volume 2003, p. 7–13 : Washington, DC.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# L'émergence de l'IA conversationnelle comme autorité cognitive : perspectives éducatives et éthiques à l'ère de Grok

Amélie Raoul
LERASS, 1186 Rte de Mende 1092, 34090 Montpellier, France
amelie.raoul@etu.univ-montp3.fr

RÉSUMÉ

Cet article analyse l'émergence d'une nouvelle autorité cognitive incarnée par les intelligences artificielles conversationnelles, en se concentrant sur le cas de Grok. Positionnée comme alternative idéologique aux modèles existants, cette IA illustre comment nous déléguons de plus en plus notre jugement à des algorithmes perçus comme objectifs ou omniscients. En interrogeant les implications éducatives, éthiques et épistémologiques de cette évolution, l'article met en lumière les risques d'une dépendance envers des systèmes dont le fonctionnement reste obscur. Il plaide en faveur d'une littératie algorithmique appliquée, d'une pédagogie de l'incertitude algorithmique et d'un outillage critique capable d'outrepasser la fascination technologique pour favoriser une réflexion autonome face aux technologies d'IA.

ABSTRACT

**The Emergence of Conversational AI as Cognitive Authority : Educational and Ethical Perspectives in the Age of Grok**

This article analyzes the emergence of a new form of cognitive authority embodied by conversational artificial intelligences, focusing on the case of Grok. Positioned as an ideological alternative to existing models, this AI illustrates how we increasingly delegate our judgment to algorithms perceived as objective or omniscient. By examining the educational, ethical, and epistemological implications of this shift, the article highlights the risks of becoming dependent on systems whose inner workings remain opaque. It advocates for an applied algorithmic literacy, a pedagogy of algorithmic uncertainty, and a critical toolkit capable of moving beyond technological fascination to foster autonomous reflection in the face of AI technologies.

MOTS-CLÉS : intelligence artificielle conversationnelle, Grok, autorité cognitive, littératie algorithmique, pédagogie critique, éthique de l'IA, incertitude algorithmique, espace public numérique, plateformes numériques, éducation critique..

KEYWORDS: conversational artificial intelligence, Grok, cognitive authority, algorithmic literacy, critical pedagogy, AI ethics, algorithmic uncertainty, digital public sphere, digital platforms, critical education..

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 70–78.

# 1 Introduction

Lancée en 2023 par xAI [1], la startup d'intelligence artificielle fondée par Elon Musk [2], Grok [3] représente une nouvelle génération d'Intelligence Artificielle (IA) conversationnelle conçue explicitement comme une alternative aux modèles dominants comme ChatGPT. Sa particularité réside dans son positionnement idéologique assumé : contrairement à d'autres IA qui affichent une neutralité de façade, Grok a été développée selon les principes défendus par Musk d'une IA « moins politiquement correcte » et « moins censurée ». Cette orientation s'inscrit dans une stratégie plus large d'influence sur le débat public concernant la régulation des intelligences artificielles et leur rôle dans la société.

Cet article propose une analyse critique de cette mutation, menée depuis une perspective ancrée en sciences de l'information et de la communication. Il mobilise une démarche interdisciplinaire croisant les apports de la sociologie des usages numériques et de la philosophie de la technique, tout en adoptant une posture critique caractéristique des SIC, attentive aux processus de médiation, aux régimes de légitimation du savoir et aux enjeux de pouvoir liés aux technologies de l'information.

L'étude de Grok présente un intérêt particulier dans le contexte actuel pour trois raisons majeures. Premièrement, son déploiement sur la plateforme X (anciennement Twitter), également propriété de Musk, lui confère une visibilité et une accessibilité sans précédent, l'intégrant directement dans un espace de débat public déjà polarisé. Deuxièmement, l'autorité symbolique dont jouit son créateur dans le champ technologique tend à conférer à cette IA une légitimité particulière auprès de certains publics. Troisièmement, la centralité des questions d'autorité cognitive et de vérité dans le projet même de Grok en fait un cas d'étude privilégié pour analyser les transformations de notre rapport au savoir à l'ère numérique.

Trois axes structurent notre réflexion : l'émergence d'une autorité cognitive automatisée, les enjeux pédagogiques associés à ce phénomène, et les pistes éducatives permettant d'en limiter les effets potentiellement délétères. Cette analyse s'inscrit dans une continuité théorique avec les travaux de (Gillespie, 2014; Eubanks, 2018; Livingstone, 2004), et prend acte des évolutions récentes en matière de littératie algorithmique (Frau-Meigs, 2024; Le Deuff & Roumanos, 2022).

# 2 Une figure d'autorité inédite

Ce phénomène manifeste l'émergence d'une nouvelle forme d'autorité cognitive, non plus fondée sur des figures incarnées (enseignants, journalistes, chercheurs), mais sur un dispositif automatisé intégré à l'espace numérique. Pour certains utilisateurs, l'IA dépasse désormais le simple rôle d'outil d'appui à la recherche pour devenir un acteur discursif autonome, un médiateur symbolique investi d'un pouvoir de validation, à la manière d'un arbitre suprême (Gillespie, 2014). Ainsi, lorsqu'un de ces utilisateurs interpelle Grok pour valider ou réfuter une information circulant en ligne, la réponse algorithmique intervient comme un verdict final. Le jugement humain se voit alors supplanté par une décision computationnelle, révélant ainsi une délégation progressive de l'esprit critique à des dispositifs automatisés (Eubanks, 2018).

Cette nouvelle pratique révèle un déplacement notable dans les mécanismes de validation des informations. On ne questionne plus, on invoque. La vérité n'est plus débattue, mais elle est délivrée. Le succès grandissant des IA génératives en général révèle une certaine fatigue cognitive face à

1. https://x.ai/
2. https://fr.wikipedia.org/wiki/Elon_Musk
3. https://grok.com/

la complexité du flot informationnel et d'une certaine injonction sociale qui pousse les individus à donner une réponse immédiate sur tout sujet. Cette dynamique s'appuie aussi sur un imaginaire collectif de la neutralité algorithmique, ce qui est particulièrement paradoxal dans le cas de Grok dont les biais sont à la fois reconnus et revendiqués. Pourtant, même avec cette transparence apparente, les processus exacts par lesquels Grok formule ses réponses restent opaques pour l'utilisateur moyen.

Le contexte spécifique de Grok, avec son positionnement « anti-woke »[4] revendiqué, crée une situation particulièrement préoccupante : une IA ouvertement non-neutre peut être perçue comme plus objective que ses concurrentes précisément parce qu'elle admet sa non-neutralité. Ce paradoxe renforce la légitimation des réponses proposées par le système et approfondit la dépendance cognitive des utilisateurs.

Il importe ici de rappeler que Grok, comme ses concurrentes, évolue rapidement. Si certaines de ses versions intègrent désormais des fonctions de recherche enrichie, avec liens vers les sources, cela ne garantit ni la qualité ni la diversité des références mobilisées. Une veille critique permanente s'impose pour évaluer l'évolution de ces dispositifs, dans la mesure où leur autorité repose autant sur des perceptions sociales que sur des capacités techniques réelles.

En définitive, la figure d'autorité que représente Grok ne se comprend ni comme neutre ni comme transparente. Elle se situe dans une zone d'ambiguïté cognitive et politique, entre performance technologique, stratégie de communication et influence idéologique. Ce constat invite à repenser de manière urgente les outils éducatifs permettant de réintroduire du doute et de la réflexivité dans l'approche des savoirs produits ou relayés par les IA conversationnelles.

# 3 Un prisme cognitif invisible

L'originalité de l'expérience Grok réside moins dans l'existence même d'une IA que dans la fonction désormais occupée par cette technologie dans l'espace public numérique. Contrairement à Google, qui propose une hiérarchie des sources, Grok présente une synthèse immédiate, et souvent univoque, sans visibilité sur ses fondements, agissant comme un prisme invisible structurant la perception du réel. Cette réponse est souvent reçue sans recul critique, car elle s'inscrit dans un contexte numérique caractérisé par l'urgence cognitive et la surcharge informationnelle (Metzger & Flanagin, 2013).

Ce phénomène s'inscrit dans un prolongement des logiques d'éditorialisation algorithmique déjà observées dans d'autres environnements numériques (Gillespie, 2014; Dominique, 2015). Mais là où les moteurs de recherche ou les fils d'actualité personnalisés suggèrent des contenus, Grok affirme directement une version synthétique d'un savoir présumé stable. L'interaction avec l'IA ne passe pas par une pluralité d'entrées, mais par une réponse unique à une requête souvent implicite, ce qui accentue le risque d'une clôture discursive précoce.

Dans cette culture numérique, les usagers cherchent des raccourcis cognitifs (Kahneman, 2011), auxquels Grok répond parfaitement par sa posture autoritaire et rassurante. Cette pratique crée un effet de clôture discursive : l'IA, avec son autorité présumée, clôt le débat avant même qu'il ne débute véritablement. Ce phénomène induit une modification structurelle majeure : le savoir cesse d'être produit par confrontation argumentée pour devenir une réponse automatique générée par un système algorithmique. Comme le montrent les travaux de (Metzger & Flanagin, 2013), les utilisateurs

4. Herrman, John. «What Does It Mean That Elon Musk's New AI Chatbot Is 'Anti-Woke'?». Intelligencer par New York Magazine. 7 novembre 2023. En ligne : https://nymag.com/intelligencer/2023/11/elon-musks-grok-ai-bot-is-anti-woke-what-does-that-mean.html [consulté le 17 avril 2025]

mobilisent des heuristiques pour juger de la crédibilité d'une information : la fluidité de l'interface, la rapidité de la réponse ou encore le style d'énonciation peuvent suffire à établir un crédit de vérité. Or, Grok répond à ces critères en s'adossant à une posture d'autorité implicite, renforcée par son ton affirmatif et par l'absence de signalement des zones d'incertitude.

Cette pratique de clôture discursive est également favorisée par la culture numérique contemporaine, qui valorise l'efficacité, la certitude et la réactivité. Les travaux de Daniel (Kahneman, 2011) sur la pensée intuitive montrent que les individus privilégient souvent des réponses rapides et intuitives, au détriment de démarches analytiques plus lentes et réflexives. Grok s'insère parfaitement dans cette logique : en offrant des réponses rapides, souvent formulées de manière assertive, elle favorise une réception passive de l'information.

Ce qui rend également le cas Grok particulièrement significatif est son intégration à la plateforme X, au sein même d'un espace de débat public numérisé. Cette insertion dans des échanges sociaux directs, notamment à travers les threads ou les réponses à des publications controversées, renforce la performativité de ses interventions. L'IA ne se contente pas de fournir une information, elle s'inscrit dans une dynamique de validation sociale, où sa réponse est immédiatement réappropriée, likée, partagée, voire instrumentalisée dans le cadre de controverses.

Cette capacité à intervenir dans des échanges interpersonnels, avec une autorité présumée mais non questionnée, place Grok dans une position inédite. L'utilisateur peut ainsi interpeller l'IA directement au sein d'une conversation ou d'un débat public, lui conférant un statut d'arbitre de la vérité dans des interactions sociales complexes. Cette contextualisation de l'autorité algorithmique dans un espace de sociabilité numérique représente une évolution majeure aux implications multiples et profondes. Ce processus renforce l'effet de naturalisation de l'autorité algorithmique, déjà identifié dans les travaux sur la médiation automatisée de l'information (Haider & Sundin, 2022).

Il convient toutefois de nuancer cette lecture. Depuis 2024, certaines versions de Grok, à l'instar de modèles comme Claude, Perplexity ou ChatGPT dans ses déclinaisons « avec navigation web », intègrent des mécanismes de référencement explicite. Des liens vers des sources sont parfois proposés, dans une logique de « recherche générative » (Butler *et al.*, 2023). Toutefois, ces liens restent souvent limités, parfois biaisés dans leur sélection, et ne permettent pas de reconstituer le cheminement exact du raisonnement algorithmique. La fonction de synthèse prime toujours sur la transparence des sources, ce qui maintient une forme de clôture épistémique.

Ainsi, Grok illustre une mutation dans la production du savoir en ligne : ce dernier tend à devenir une réponse plutôt qu'un processus, une évidence plutôt qu'une élaboration, une certitude immédiate plutôt qu'un objet de discussion. Cette évolution appelle une révision profonde des cadres éducatifs, notamment en matière de littératie algorithmique. Comprendre le rôle des IA dans la structuration cognitive du réel devient une compétence essentielle, qui dépasse largement la seule maîtrise technique des outils pour embrasser les enjeux sociopolitiques de l'information numérique.

# 4 Enjeux éducatifs et éthiques

Ces évolutions que nous avons évoquées plus haut posent un défi crucial aux éducateurs. Si l'acte même de douter se délègue à une IA, comment préserver chez les jeunes la capacité critique d'interroger méthodologiquement et épistémologiquement l'information ? Comme le souligne (Livingstone, 2004), les compétences numériques dépassent largement la maîtrise technique et incluent l'aptitude à analyser les dispositifs sociotechniques qui régulent l'accès à l'information.

Former à la littératie informationnelle aujourd'hui signifie enseigner aux apprenants à poser les bonnes questions à une IA, à trianguler les sources, à contextualiser et argumenter face aux réponses reçues (Zawacki-Richter *et al.*, 2019). Ce processus ne relève pas uniquement d'une instruction méthodologique ; il implique une acculturation critique aux logiques techniques et éditoriales des agents conversationnels. Cette formation implique une évolution des méthodologies pédagogiques vers une pédagogie du doute et de la métacognition, capable de décrypter les biais implicites des algorithmes. À titre d'exemple, une expérimentation menée en 2024 dans une université européenne montre que les étudiants formés spécifiquement à l'usage critique des IA conversationnelles développent une meilleure compréhension des biais algorithmiques et un esprit critique renforcé face aux réponses automatisées (Karsenti, 2018). Les travaux de (Frau-Meigs, 2024) apportent également une perspective pertinente en mettant l'accent sur la résilience face à la désinformation à travers la littératie médiatique et informationnelle. Leur étude explore comment des programmes éducatifs peuvent efficacement aider les étudiants à déceler les fausses informations, ce qui est particulièrement pertinent dans le contexte des IA génératives où la véracité et la crédibilité des sources sont souvent remises en question.

Dans le cas spécifique de Grok, l'enjeu éducatif est double. D'une part, il s'agit de sensibiliser les apprenants aux biais explicites revendiqués par ses concepteurs. D'autre part, il faut développer chez eux la capacité à identifier les biais implicites qui, précisément parce que certains biais sont admis ouvertement, peuvent passer inaperçus. Cette situation inédite réclame le développement d'une méta-critique des IA : apprendre à discerner non seulement les biais d'un système, mais aussi la manière dont la présentation de ces biais peut elle-même constituer une stratégie rhétorique.

Ce type de réflexivité suppose une approche pédagogique différenciée selon les publics. Dans le secondaire, par exemple, des dispositifs d'analyse comparative entre moteurs de recherche, IA conversationnelles et encyclopédies numériques peuvent initier une première prise de distance critique. Dans l'enseignement supérieur, des études de cas problématisées autour de réponses contradictoires fournies par différentes IA permettent d'introduire les notions d'autorité, de sourcing et de conflictualité informationnelle. Cette contextualisation pédagogique apparaît essentielle pour intégrer pleinement les enjeux de l'IA dans une perspective éducative.

Enfin, une réflexion éthique s'impose. L'introduction massive des IA dans les contextes d'apprentissage soulève la question du coût écologique et infrastructurel d'un recours systématisé à ces outils. Si l'on encourage les usages critiques des IA dans les formations, encore faut-il s'interroger sur leur soutenabilité, tant du point de vue énergétique que du point de vue des inégalités d'accès. Ces considérations doivent désormais être articulées aux projets pédagogiques, notamment dans le cadre d'une éducation au numérique responsable.

# 5 Pour une littératie algorithmique appliquée

Dans le but précédemment évoqué, le concept de littératie algorithmique constitue un fondement essentiel pour développer une posture critique face aux systèmes tels que Grok (Le Deuff & Roumanos, 2022). Cette compétence transcende la simple analyse de contenus pour englober la compréhension des mécanismes algorithmiques sous-jacents, leurs contraintes inhérentes, leurs paradigmes conceptuels et leurs implications socio-cognitives. Des recherches récentes (Haider & Sundin, 2022) soulignent aussi la pertinence de la translittératie, notion complémentaire, définie comme l'aptitude à naviguer entre différents écosystèmes informationnels en modulant ses stratégies cognitives. Ces compétences deviennent cruciales face à des agents conversationnels qui agrègent l'information sans

en expliciter les sources. Elles permettent également de situer le fonctionnement de l'IA dans un continuum de pratiques médiatiques, entre automatisation de la hiérarchisation de l'information et invisibilisation des logiques éditoriales. Former à la littératie algorithmique appliquée consiste donc à doter les individus de la capacité à interroger la validité généralisée d'une réponse, à repérer les simplifications excessives et à resituer l'information dans des contextes plus vastes. De ce point de vue, la littératie algorithmique relève à la fois de l'analyse critique et de l'auto-régulation cognitive.

Cela implique une redéfinition des apprentissages numériques, qui doivent aller au-delà des compétences procédurales. Il ne s'agit plus simplement d'apprendre à manipuler une intelligence artificielle, mais de développer la capacité à en analyser les logiques sous-jacentes, à décrypter les intentions implicites, les arbitrages de conception et les modèles cognitifs intégrés.

L'enjeu éthique revêt ici une importance fondamentale : il s'agit de former des citoyens aptes à dissocier autorité algorithmique et vérité, et à cultiver une autonomie cognitive ancrée dans le doute méthodique. Cette orientation s'aligne d'ailleurs avec les recommandations de l'UNESCO (UNESCO, 2023) pour une éducation critique aux médias et à l'information à l'ère des intelligences artificielles. Celles-ci mettent en évidence la nécessité d'intégrer ces compétences dans les cursus tout au long de la vie, depuis l'enseignement secondaire jusqu'à la formation continue. Dans un environnement pédagogique, cette approche nécessite l'élaboration de dispositifs d'apprentissage hybrides où les apprenants sont exposés à des études de cas concrets d'interaction avec diverses intelligences artificielles conversationnelles. L'usage d'ateliers d'analyse critique, de jeux de rôle ou de débats argumentés autour de réponses divergentes entre IA permet de matérialiser les biais et d'entraîner à la vérification des contenus. De tels dispositifs permettent non seulement de rendre visibles les mécanismes algorithmiques mais aussi de revaloriser le rôle de l'incertitude comme moteur du raisonnement critique.

# 6 Vers une pédagogie de l'incertitude algorithmique

Pour répondre aux défis posés par l'émergence de cette nouvelle autorité cognitive, nous proposons le développement d'une « pédagogie de l'incertitude algorithmique » particulièrement adaptée au cas de Grok. Cette approche vise à transformer l'incertitude inhérente aux systèmes d'IA en ressources pédagogiques plutôt qu'en obstacles à surmonter.

Cette pédagogie s'articule autour de trois principes applicables immédiatement dans les contextes éducatifs :

1. La confrontation systématique : Encourager les apprenants à soumettre une même question à différentes IA (Grok, mais aussi ses concurrents) pour comparer leurs réponses et analyser les divergences. Cette méthode simple permet de révéler concrètement la non-neutralité des algorithmes et d'initier une réflexion critique sur leurs présupposés.
2. L'analyse des sources dissimulées : Développer des exercices où les apprenants tentent de reconstituer les sources probables mobilisées par l'IA pour construire sa réponse. Ce travail d'archéologie informationnelle renforce la conscience des processus de médiation à l'œuvre dans la production algorithmique du savoir.
3. La réhabilitation de l'incertitude : Valoriser explicitement dans l'évaluation scolaire la capacité à formuler des doutes méthodiques et à suspendre son jugement face à des questions complexes, plutôt que de privilégier systématiquement les réponses assertives que proposent les IA comme Grok.

Ces approches pourraient contribuer à développer chez l'usager une relation plus critique et réflexive aux IA conversationnelles. Cette hypothèse s'appuie sur des travaux antérieurs dans le domaine de l'éducation aux médias numériques. En effet, (Buckingham, 2019) a démontré que l'exposition des élèves à des informations contradictoires renforce leur capacité à exercer un jugement critique. Ainsi, une pédagogie de l'incertitude algorithmique pourrait être particulièrement pertinente face à des technologies comme Grok.

À cela s'ajoute une exigence méthodologique : il ne s'agit pas seulement d'entraîner au doute, mais de structurer des cadres d'analyse permettant de questionner la forme même des réponses produites. Les enseignants peuvent, par exemple, intégrer des grilles de lecture inspirées des travaux en didactique de l'information, afin de faire émerger les implicites narratifs, les effets d'autorité syntaxique et les procédés de clôture discursive utilisés par l'IA. Ce type de décodage s'avère particulièrement utile dans les formations supérieures, où l'enjeu n'est plus seulement de comprendre l'IA, mais de produire un savoir informé sur son fonctionnement.

Par ailleurs, cette pédagogie n'est pas réservée à l'enseignement formel. Elle peut s'inscrire dans des initiatives de médiation scientifique, de formation professionnelle ou de sensibilisation citoyenne. Des campagnes éducatives à destination des parents, des salariés ou des publics éloignés du numérique permettraient d'élargir le champ d'action de cette approche critique. Ainsi, divers dispositifs pourraient être mis en œuvre, tant dans un cadre scolaire que chez l'adulte (campagnes de sensibilisation, formations au sein des entreprises, etc.)

# 7 Conclusion

La tendance grandissante des usagers à considérer que Grok incarne la vérité, telle une autorité omnisciente, témoigne d'une fascination pour les IA, qui n'est pas nouvelle mais prend aujourd'hui des dimensions et des aspects inédits. Le modèle conversationnel s'est tellement intégré aux usages qu'il est devenu naturel de lui déléguer son discernement. Or, cette banalisation est précisément le danger : si le recours à Grok devient une évidence pour les utilisateurs, l'incidence du jugement humain risque de drastiquement diminuer dans de nombreux aspects de notre quotidien, si bien que les conséquences de cette tendance seraient quasi impossibles à anticiper. La production du savoir ne doit pas devenir une interaction individuelle sans recul critique avec une interface opaque, mais rester un processus collectif critique et réflexif (Floridi, 2019). Face à cette IA dont la particularité est d'assumer explicitement une forme de subjectivité algorithmique, une réponse éducative appropriée devrait permettre aux utilisateurs de développer non seulement des compétences critiques générales, mais aussi une compréhension fine des dynamiques de pouvoir et d'influence qui traversent le champ technologique.

Cette reconfiguration de l'autorité cognitive s'inscrit dans une transformation plus large des régimes de vérité contemporains. Comme le montre l'étude menée par (Butler *et al.*, 2023), une part significative des jeunes utilisateurs de plateformes sociales exprime une confiance croissante envers les IA génératives pour obtenir des informations, parfois même plus élevée que celle accordée aux médias traditionnels. Ce glissement perceptif, s'il reste à stabiliser empiriquement sur le long terme, mérite une attention soutenue de la recherche en sciences de l'information et de la communication.

À mesure que les intelligences artificielles conversationnelles acquièrent un statut d'interlocuteurs légitimes au sein de l'espace public numérique, une interrogation essentielle émerge : sous quelles conditions sommes-nous prêts à confier notre faculté de jugement à ces systèmes algorithmiques ? L'exemple de Grok, tant représentatif qu'unique en son genre, nous conduit à examiner de façon

critique les manifestations contemporaines de l'autorité cognitive ainsi que les processus d'élaboration des connaissances dans notre ère numérique. Il ouvre également un chantier de recherche interdisciplinaire sur les formes émergentes de légitimation du savoir algorithmique, qui mobilise à la fois les outils des sciences sociales, de l'éducation critique, de l'éthique appliquée et de la philosophie politique. La pédagogie de l'incertitude algorithmique n'est pas un simple correctif temporaire, mais un levier conceptuel et méthodologique pour réinscrire le doute, la pluralité et la réflexivité au cœur de notre rapport aux technologies intelligentes. Plus que jamais, la formation au numérique implique une sensibilisation aux enjeux économiques, politiques et éthiques des technologies d'IA, au-delà de leurs seules dimensions techniques.

# Références


BUCKINGHAM D. (2019). *The media education manifesto*. John Wiley & Sons.

BUTLER J., JAFFE S., BAYM N., CZERWINSKI M., IQBAL S., NOWAK K., RINTEL S., SELLEN A., VORVOREANU M., ABDULHAMID N. G., AMORES J., ANDERSEN R., AWORI K., AXMED M., BOYD D., BRAND J., BUSCHER G., CARIGNAN D., CHAN M., COLEMAN A., COUNTS S., DAEPP M., FOURNEY A., GOLDSTEIN D. G., GORDON A., HALFAKER A. L., HERNANDEZ J., HOFMAN J., LAY-FLURRIE J., LIAO V., LINDLEY S., MANIVANNAN S., MCILWAIN C., NEPAL S., NEVILLE J., NYAIRO S., O'NEILL J., POZNANSKI V., RAMOS G., RANGAN N., ROSEDALE L., ROTHSCHILD D., SAFAVI T., SARKAR A., SCOTT A., SHAH C., SHAH N. P., SHAPIRO T., SHAW R., SIMKUTE A., SUH J., SURI S., TANASE I., TANKELEVITCH L., TROY A., WAN M., WHITE R. W., YANG L., HECHT B. & TEEVAN J. (2023). *Microsoft New Future of Work Report 2023*. Rapport interne MSR-TR-2023-34, Microsoft.

DOMINIQUE C. (2015). À quoi rêvent les algorithmes. nos vies à l'heure des big data. *Paris, Éditions du Seuil & La république des idées*.

EUBANKS V. (2018). *Automating inequality : How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

FLORIDI L. (2019). *The logic of information : A theory of philosophy as conceptual design*. Oxford University Press.

FRAU-MEIGS D. (2024). Algorithm literacy as a subset of media and information literacy : Competences and design considerations. *Digital*, **4**(2), 512–528. DOI : 10.3390/digital4020026.

GILLESPIE T. (2014). The relevance of algorithms. *Media technologies : Essays on communication, materiality, and society*, **167**(2014), 167.

HAIDER J. & SUNDIN O. (2022). *Paradoxes of media and information literacy : The crisis of information*. Taylor & Francis. DOI : 10.4324/9781003163237.

KAHNEMAN D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux*.

KARSENTI T. (2018). Intelligence artificielle en éducation : L'urgence de préparer les futurs enseignants aujourd'hui pour l'école de demain. *Formation et profession*, **26**(3), 112–119.

LE DEUFF O. & ROUMANOS R. (2022). Enjeux définitionnels et scientifiques de la littératie algorithmique : entre mécanologie et rétro-ingénierie documentaire. *tic&société*, **15**(2-3| 2ème semestre 2021-1er semestre 2022), 325–360. DOI : 10.4000/ticetsociete.7105.

LIVINGSTONE S. (2004). Media literacy and the challenge of new information and communication technologies. *The communication review*, **7**(1), 3–14.

METZGER M. J. & FLANAGIN A. J. (2013). Credibility and trust of information in online environments : The use of cognitive heuristics. *Journal of Pragmatics*, **59**, 210–220. Biases and constraints in communication : Argumentation, persuasion and manipulation, DOI : https ://doi.org/10.1016/j.pragma.2013.07.012.

UNESCO (2023). *Guidelines for the governance of digital platforms : safeguarding freedom of expression and access to information through a multi-stakeholder approach*. Paris, France : UNESCO. Foreword by the Director-General of UNESCO, Audrey Azoulay. Includes bibliography. Available in multiple languages. Licensed under CC BY-SA 3.0 IGO.

ZAWACKI-RICHTER O., MARÍN V. I., BOND M. & GOUVERNEUR F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators ? *International journal of educational technology in higher education*, **16**(1), 1–27. DOI : 10.1186/s41239-019-0171-0.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# MALIN : MAnuels scoLaires INclusifs

Elise Lincker[1] Léa Pacini[1,2] Mohamed Amine Lasheb[1]
Olivier Pons[1] Jérôme Dupire[1] Camille Guinaudeau[3] Céline Hudelot[4]
Vincent Mousseau[4] Isabelle Barbet[1] Caroline Huron[2,5]
(1) Cedric, CNAM, Paris, France
(2) SEED, Inserm, Paris, France
(3) LISN, Université Paris-Saclay, CNRS, Orsay, France
(4) MICS, CentraleSupélec, Université Paris-Saclay, Orsay, France
(5) Learning Planet Institute, Paris, France
elise.lincker@lecnam.net, lea.pacini@lecnam.net,
mohamed-amine.lasheb@lecnam.net, olivier.pons@lecnam.net,
jerome.dupire@lecnam.net, guinaudeau@universite-paris-saclay.fr,
celine.hudelot@centralesupelec.fr, vincent.mousseau@centralesupelec.fr,
isabelle.barbet@lecnam.net, caroline.huron@learningplanetinstitute.org

RÉSUMÉ

L'accès aux manuels scolaires constitue un enjeu majeur pour l'éducation inclusive. Le projet MALIN vise à automatiser l'adaptation des manuels scolaires, convertissant un manuel numérique au format PDF en une version structurée, puis en une version accessible. Le projet cible en priorité la dyspraxie, tout en posant les bases d'adaptations pour d'autres handicaps. Les adaptations proposées, conçues avec des experts de l'accessibilité, sont évaluées auprès d'enfants en situation de handicap.

ABSTRACT

**Inclusive Textbooks**

Ensuring access to textbooks is crucial for inclusive education. The MALIN project aims to automate textbook adaptation from PDF files by converting them into structured and accessible versions. The project focuses on Developmental Coordination Disorder, while also offering solutions that can benefit learners with other disabilities. The adaptations are designed in collaboration with experts and evaluated with the target audience.

MOTS-CLÉS : Education inclusive, manuels scolaires, dyspraxie.

KEYWORDS: Inclusive education, textbooks, developmental coordination disorder.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

## 1 Contexte

Les manuels scolaires sont un outil pédagogique essentiel, quasi systématiquement utilisé à l'école en France. Pourtant, leurs versions numériques ne sont pas accessibles aux élèves en situation de handicap, freinant leur apprentissage. Le projet MALIN vise à automatiser l'adaptation de ces manuels pour les rendre accessibles. Nous concevons des adaptations et développons une chaîne de traitement

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 79–82.

permettant de convertir un manuel au format PDF en une version adaptée. Un manuel adapté conserve le même contenu et les mêmes intentions pédagogiques que le manuel original, mais propose un mode d'interaction compatible avec les besoins de l'élève. Nous ciblons la dyspraxie : pour les élèves dyspraxiques, les activités sont rendues interactives pour compenser les difficultés d'écriture manuscrite, et la présentation est adaptée pour prendre en compte les troubles d'organisation du regard. Ces adaptations résolvent les problématiques liées au geste moteur, à l'écriture manuscrite et à l'organisation visuospatiale. Elles peuvent également bénéficier aux élèves présentant des troubles moteurs, des troubles de l'attention, des troubles du spectre de l'autisme ou encore une déficience visuelle.

# 2 Méthodologie

## 2.1 Conception et évaluation des adaptations

Les adaptations sont conçues par l'association Cartable Fantastique [1] en croisant le regard de chercheurs en sciences cognitives spécialistes de la dyspraxie, d'enseignants habitués à adapter pour inclure et de personnes dyspraxiques. Elles sont centrées sur les besoins des utilisateurs finaux et tiennent compte des difficultés d'écriture manuscrite et d'organisation du regard de ces enfants. Par exemple, un exercice de type « Recopie les phrases et souligne les verbes » est remplacé par un exercice numérique de type « Dans ces phrases, clique sur les verbes ».

Une étude en Single Case Experimental Design est actuellement menée auprès d'enfants dyspraxiques. Cette étude a pour objectif de comparer différentes présentations d'exercices (présentation typique de manuel scolaire vs deux présentations type Cartable Fantastique) afin de déterminer laquelle ou lesquelles amélioreront les performances des enfants en termes de temps de réaction et de nombre d'erreurs (Pacini *et al.*, 2023).

## 2.2 Automatisation

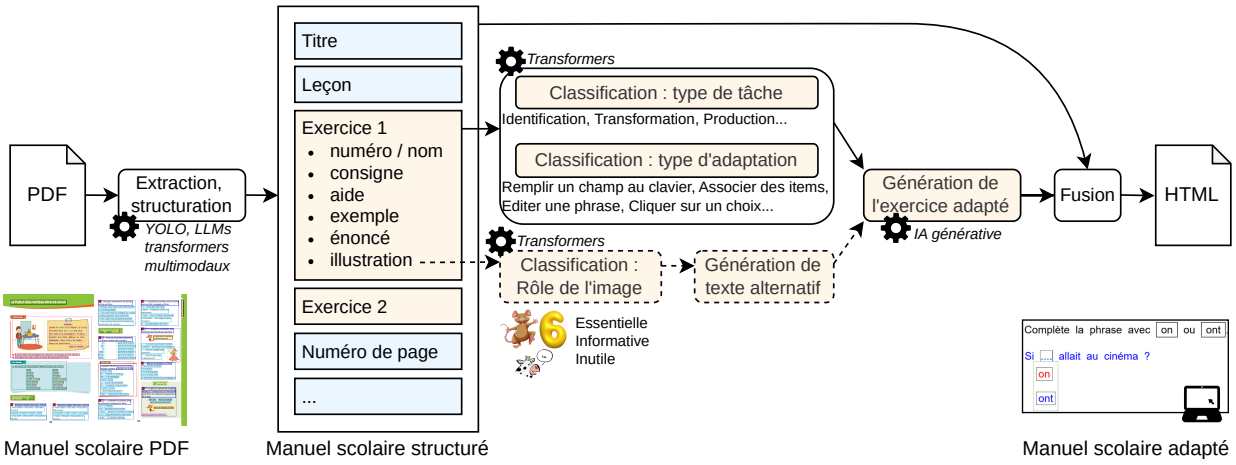



FIGURE 1 – Chaîne de traitement : du manuel PDF au manuel adapté

1. https://www.cartablefantastique.fr/

La chaîne de traitement développée est illustrée en Figure 1. Elle permet de convertir un manuel scolaire au format PDF en une version accessible. Elle se compose de plusieurs étapes :

**1. Structuration du contenu.**
Nous avons défini un modèle de manuel scolaire, permettant une représentation normalisée et structurée des contenus à l'échelle du document et de la page (Lincker *et al.*, 2023b). L'étape d'extraction et de structuration automatique du contenu repose sur une approche textuelle (Lincker *et al.*, 2023b) ou visuelle (Lasheb *et al.*, 2025).

**2. Classification des exercices.**
Les exercices extraits sont ensuite classés à l'aide de modèles transformers multimodaux. La classification s'effectue selon deux axes : la nature de la tâche pédagogique sollicitée, et le type d'adaptation appliqué dans le cadre de la dyspraxie (Lincker *et al.*, 2023a).

**3. Gestion des images.**
Lorsque les activités pédagogiques contiennent des illustrations, celles-ci sont classées selon leur fonction (essentielle, informative, décorative) (Yadav *et al.*, 2024). Cette classification détermine si l'image est conservée dans l'adaptation, et guide la génération de texte alternatif.

**4. Génération des adaptations.**
Enfin, les exercices sont transformés en fichiers HTML interactifs, tenant compte des déficits moteurs, des troubles de l'écriture manuscrite et des aspects de présentation visuelle. Pour garantir la conformité aux originaux et l'accessibilité des contenus, une vérification manuelle systématique reste indispensable.

## 2.3 Limites

En l'absence de jeux de données librement exploitables, nous n'avons accès qu'à un nombre restreint de données utilisables. Une exception aux droits d'auteur permet l'adaptation des ouvrages pour les personnes en situation de handicap. Dans le cadre de cette exception, l'association Cartable Fantastique est agréée pour pouvoir adapter les manuels scolaires. Toutefois, les corpus annotés que nous avons constitués ne peuvent pas être partagés en raison de restrictions liées aux droits d'auteur.

Par ailleurs, nos méthodes visent à adapter des modèles existants à un contexte spécifique : des données scolaires francophones, en faible quantité, bruitées et déséquilibrées.

## 2.4 Ouverture

Afin d'accélérer l'adaptation des manuels scolaires, nous développons une plateforme numérique utilisable par les personnes en charge de l'adaptation. Elle intégrera l'ensemble des étapes automatiques du processus tout en permettant aux utilisateurs d'intervenir à chaque phase pour corriger, valider ou modifier les adaptations générées, et assurer un contrôle qualité. La plateforme est conçue en co-construction avec les utilisateurs finaux.

Enfin, si nos premiers travaux ciblent la dyspraxie, les adaptations proposées bénéficieront plus largement à tous les élèves qui ont des besoins de compensation de difficultés visuelles ou motrices. De plus, la structuration fine du contenu permet d'envisager des adaptations alternatives au HTML, notamment pour les braillistes ou compatibles avec les lecteurs d'écran.

## Remerciements

Le projet MALIN a reçu un soutien de l'Agence Nationale de la Recherche (convention ANR-21-CE38-0014). Ce travail a été financé par le LISN.

## Références

LASHEB M. A., PONS O., BEKKOUCHE M., LINCKER E., BARBET I. & HURON C. (2025). Extracting and structuring textbooks for inclusive education : A computer vision approach. In *Proceedings of the 25th IEEE International Conference on Advanced Learning Technologies (ICALT 2025)*. A paraître.

LINCKER E., GUINAUDEAU C., PONS O., DUPIRE J., HUDELOT C., MOUSSEAU V., BARBET I. & HURON C. (2023a). Noisy and unbalanced multimodal document classification : Textbook exercises as a use case. In *Proceedings of the 20th International Conference on Content-based Multimedia Indexing (CBMI 2023)*, Orléans, France.

LINCKER E., PONS O., GUINAUDEAU C., BARBET I., DUPIRE J., HUDELOT C., MOUSSEAU V. & HURON C. (2023b). Layout and activity-based textbook modeling for automatic pdf textbook extraction. In *Proceedings of the 5th International Workshop on Intelligent Textbooks, 24th International Conference on Artificial Intelligence in Education (AIED 2023)*, Tokyo, Japan.

PACINI L., DUPIRE J., BARBET I., PONS O., GUINAUDEAU C., MOUSSEAU V., HUDELOT C. & HURON C. (2023). Textbook's accessibility for children with dyspraxia and visual disability. In *17th International Conference of the Association for the Advancement of Assistive Technology in Europe (AAATE 2023)*.

YADAV S., LINCKER E., HURON C., MARTIN S., GUINAUDEAU C., SATOH S. & SHUKLA J. (2024). Towards inclusive education : Multimodal classification of textbook images for accessibility. In *Proceedings of the 31st Conference on Multimedia Modeling (MMM 2025)*.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Profilage comportemental dans les jeux vidéo éducatifs via des réseaux convolutifs graphiques : le cas de GraphoGameFrançais

Emna Ammari[1] Patrice Bellot[1] Ambre Denis-Noël[2] Johannes C.Ziegler[2]
(1) Aix Marseille Univ, CNRS, LIS, Marseille, France
(2) Aix Marseille Univ, CNRS, CRPN, Marseille, France
emna.ammari@lis-lab.fr, patrice.bellot@univ-amu.fr,
johannes.ziegler@univ-amu.fr, ambre.denis@univ-amu.fr

## RÉSUMÉ

Les données comportementales des jeux vidéo ainsi que les traces de joueurs suscitent un intérêt croissant, tant pour la recherche que pour l'industrie du jeu. Ces données peuvent notamment enrichir l'expérience de jeu et améliorer l'identification automatique des profils des joueurs. Dans cet article, nous nous intéressons principalement aux données du jeu sérieux GraphoGame, un outil innovant d'aide à l'apprentissage de la lecture, offrant un environnement interactif pour les apprenants. Nous cherchons notamment à évaluer l'impact de ce jeu sur la performance des élèves en lecture via le profilage comportemental des joueurs et un apprentissage à base de graphes. Ainsi, deux techniques d'intégration basées sur des réseaux convolutifs, GraphSAGE et ECCConv, sont mises à profit pour classifier les graphes d'interactions des joueurs. Les résultats montrent qu'ECCConv surpasse GraphSAGE, mais que leurs prédictions combinées peuvent améliorer la classification, confirmant l'impact éducatif de GraphoGame même chez les élèves les plus avancés.

## ABSTRACT

**Behavioral profiling in educational video games through graph convolutional networks : GraphoGameFrançais use case.**

Recently, behavioral data of computer games and players actions have turned to be a pivotal interest for both researchers and game industry. These data could substantially enhance the gaming experience and machine learning tasks to build more actionable groups of players. In this paper, we're more interested in a serious learning game data GraphoGame, an advanced tool providing learners with a dynamic educational environment. In light of this, the study focuses on revealing the true impact of GraphoGame on students' reading performance through behavior profiling and graph representation learning. Accordingly, two convolutional network-based embedding techniques, GraphSAGE and ECCConv, are leveraged to classify player interactions graphs. The results show that ECCConv can outperform GraphSAGE, yet their combined predictions enhance classification, underlining the educational impact of GraphoGame even among advanced students.

MOTS-CLÉS : Jeux vidéo sérieux – profilage des joueurs – réseaux convolutifs de graphes – intégrations de graphes.

KEYWORDS: Serious video games – players profiling – graph convolutional networks – graph embeddings.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 83–93.

# 1 Introduction

En quelques décennies, les jeux vidéo se sont imposés comme une source majeure de divertissement et de plaisir au niveau mondial. Devenu une tendance en plein essor, ce domaine attire des joueurs de tous âges et de tous horizons. Certains jeux se distinguent par un objectif « sérieux » et sont appréciés pour leur impact éducatif (Wong *et al.*, 2007). Ces jeux proposent une approche interactive et diversifiée, à même de répondre à des objectifs de formation, de communication ou de sensibilisation plutôt que de simplement divertir. En France, l'industrie du jeu éducatif affiche une progression spectaculaire, entraînant plusieurs startups à cibler en priorité les établissements scolaires et les centres de formation. Selon (Lee & Peng, 2006), les jeux éducatifs s'adressent principalement aux enfants de moins de dix ans et visent à renforcer leurs compétences en lecture ainsi que leur expérience éducative enrichie par des activités ludiques. Ce type de jeux constitue un vecteur privilégié favorisant un apprentissage actif et contextualisé ainsi que l'engagement scolaire des élèves.

Dans le but d'assurer une expérience utilisateur enrichissante, les concepteurs de jeux manifestent un intérêt croissant pour l'analyse en temps réel des interactions des joueurs. Ces observations obtenues durant le *gameplay* améliorent la qualité des jeux et permettent une meilleure compréhension de l'implication des joueurs, tout en maintenant leur intérêt. Par conséquent, les indicateurs de *gameplay*, définis comme les traces numériques laissées par les joueurs (Drachen *et al.*, 2013), constituent des éléments majeurs dans les processus de modélisation comportementale des joueurs.

Ce travail s'inscrit autour du jeu sérieux GraphoGameFrançais relevant d'une dynamique interdisciplinaire, combinant l'informatique, les sciences de l'éducation et la psychologie cognitive. Il est basé sur une version française du jeu finlandais GraphoGame, un outil numérique d'aide à l'apprentissage de la lecture. La version française de GraphoGame (GGF) [1] propose un programme d'entraînement audiovisuel visant à renforcer les compétences fondamentales en lecture chez les enfants d'âge préscolaire et primaire, y compris ceux présentant des troubles spécifiques des apprentissages.

# 2 Analyse d'impact des jeux sérieux et identification de profils

Bien que de nombreuses recherches antérieures aient mis en lumière l'efficacité des jeux vidéo sérieux à visée éducative (Wong *et al.*, 2007; Smith, 2008; Laamarti *et al.*, 2014), la plupart se limitent à des revues de la littérature ou des analyses conceptuelles. En revanche, les recherches empiriques fondées sur des données issues du gameplay réel demeurent rares.

Deux études ciblées (Lassault *et al.*, 2022; Ruiz *et al.*, 2017) ont été menées afin d'évaluer l'efficacité de GGF en comparaison avec un autre outil d'entraînement des compétences mathématiques, FieteMath [2]. Bien que ces travaux aient confirmé l'apport distinct de GGF auprès d'enfants à risque de dyslexie scolarisés, leurs analyses reposaient exclusivement sur des données et des évaluations des activités de lecture en dehors du jeu ou sur quelques indicateurs tels que le pourcentage de bonnes réponses ou le nombre de niveaux complétés. Ainsi, les données séquentielles des traces d'interaction enregistrées durant le *gameplay* ont été exclues de ces analyses.

En outre, les jeux vidéo commerciaux non éducatifs ont fait l'objet de nombreuses analyses, reposant sur divers attributs et méthodes. Des caractéristiques dérivées à l'aide d'outils tels que des caméras

1. https://graphogame.com/le-jeu/
2. https://www.ahoiii.com

(Kim *et al.*, 2015) ou des capteurs enregistrant les réactions des joueurs (Smerdov *et al.*, 2023) ont été exploitées pour identifier des clusters de comportements distincts, comme le souligne également la revue de Hooshyar *et al.* (2018). D'autres études, en revanche, se sont appuyées sur les fichiers logs de *gameplay* comme source principale de données d'interactions (Sifa *et al.*, 2014; Saas *et al.*, 2016; Menéndez *et al.*, 2014), en recourant à des méthodes non supervisées pour l'analyse des parcours des joueurs. Par ailleurs, plusieurs approches fondées sur des réseaux de neurones, telles que les architectures à mémoire longue à court terme LSTM (Shahzad Farooq *et al.*, 2021), les modèles d'apprentissage par renforcement profond (Gharbi *et al.*, 2024), et, plus récemment, le modèle *Actionable Forecasting Network* (Jagirdar *et al.*, 2024), ont été mobilisées pour l'identification des profils d'utilisateurs. Les représentations vectorielles des interactions des joueurs se sont également révélées pertinentes pour la modélisation des comportements complexes, notamment à travers l'usage de techniques d'intégration de nœuds (*node embedding*) telles que Node2Vec et *Large-scale Information Network Embedding* (Shah & Thue, 2023), ou encore des encodeurs basés sur des réseaux de neurones (Sapienza *et al.*, 2019) et des modèles de langue (Wang *et al.*, 2024). Enfin, les réseaux de neurones graphiques (*Graph Neural Networks*, GNNs) se sont révélés particulièrement efficaces dans l'analyse du *gameplay*. Cependant, l'attention s'est majoritairement portée sur des représentations vectorielles au niveau des nœuds, en association avec d'autres méthodes d'apprentissage automatique ou techniques de partitionnement telles que les k-moyennes (Melo *et al.*, 2020).

En complément de ces études, notre travail s'inscrit dans une direction peu explorée, en mettant l'accent sur les plongements (*embeddings*) au niveau des graphes pour la modélisation des joueurs. L'ensemble du graphe d'interaction est encodé dans un espace latent, permettant un profilage direct, sans avoir recours à des classificateurs supplémentaires. Sur le plan applicatif, cet article a pour objectif de mettre en évidence et de valider l'impact éducatif du GGF sur l'amélioration des compétences en lecture, en exploitant les traces comportementales générées en jeu pour une caractérisation fine de profils d'apprenants à l'aide de techniques d'apprentissage automatique sur graphes.

# 3 GraphoGame : expérimentation et données

GraphoGame constitue un cadre méthodologique innovant favorisant une intégration effective des technologies numériques au sein des environnements d'apprentissage scolaires. Une expérimentation a été conduite auprès de plusieurs groupes d'élèves de niveau CP[3], durant des cours de français durant l'année scolaire 2017–2018. Les sessions d'entraînement sur tablette étaient organisées à raison de quatre par semaine, chacune durant entre 15 et 20 minutes, selon le rythme et le taux d'erreurs de chaque élève. Au total, 451 élèves âgés de 5 à 8 ans ont pris part à l'expérimentation. Ces élèves ont ainsi été engagés, via l'application, dans une variété d'activités ludiques de lecture, conçues pour renforcer leurs compétences en décodage de lettres et de mots ainsi que le développement de la fluidité en lecture. Ces activités s'inscrivent dans une progression pédagogique structurée en 67 unités (séquences), chacune centrée sur un contenu spécifique (Lassault *et al.*, 2022).

À titre d'exemple, la première unité introduit les voyelles orales simples (ex. : 'a', 'i', 'e', 'o'), les consonnes continues (ex. : 'f', 'j', 'l', 'r') et des mots monosyllabiques (ex. : 'ou', 'la', 'sol', 'four'). La quatrième, quant à elle, porte sur les voyelles nasales (ex. : 'an', 'on', 'in', 'un') tandis que la troisième propose des exercices de discrimination visuelle (ex. : 'u' vs 'n'), phonémique (ex. : 't' vs 'd') et combinée (ex. : 'p' vs 'b').

3. CP : Cours préparatoire, première année de l'école élémentaire

Chaque unité pédagogique se compose d'environ 10 niveaux, proposant un exercice ludique distinct. La durée moyenne d'un niveau est de 2 à 3 minutes et comprend généralement entre 10 et 15 essais. Selon l'exercice, l'enfant, muni d'un casque, doit associer un stimulus auditif à l'élément visuel correspondant affiché à l'écran. Si le taux d'erreur dépasse 25 %, le niveau est automatiquement réinitialisé. Après cinq tentatives, l'élève peut néanmoins accéder au niveau suivant. L'application permet également à l'élève de revenir sur les niveaux non réussis ou d'ignorer ceux déjà maîtrisés.

Les interactions des élèves avec GraphoGameFrançais génèrent une séquence horodatée d'événements. Ces données, issues de 1 548 100 actions consignées, constituent la principale source des métriques de *gameplay* analysées dans le cadre de cet article, à partir desquelles quatre indicateurs clés sont sélectionnés (table 1).

| Indicateur | Description |
|---|---|
| accuracy | Indique si la réponse du joueur est correcte (1) ou non (0) |
| RTclean | Temps de réaction du joueur pour donner des réponses correctes |
| nRepLevel | Nombre de répétitions d'un niveau donné par un joueur |
| abortedLevel | Indique si le joueur a quitté le niveau avant de le terminer (1) ou non (0) |

TABLE 1 – Indicateurs comportementaux du *gameplay*.

À la suite du prétraitement des données, 440 joueurs ont été retenus pour la modélisation, soit 440 séquences de données détaillant l'intégralité de leurs actions dans le jeu. Les participants ont complété 49 unités pédagogiques, les premières étant les plus fréquemment jouées, comme l'illustre la figure 1. Au total, 579 niveaux ont été explorés. Toutefois, certaines des unités n'ont pas été entièrement parcourues par l'ensemble des élèves. L'unité pédagogique 4.1 représente, pour la majorité d'entre eux, la phase la plus avancée atteinte.

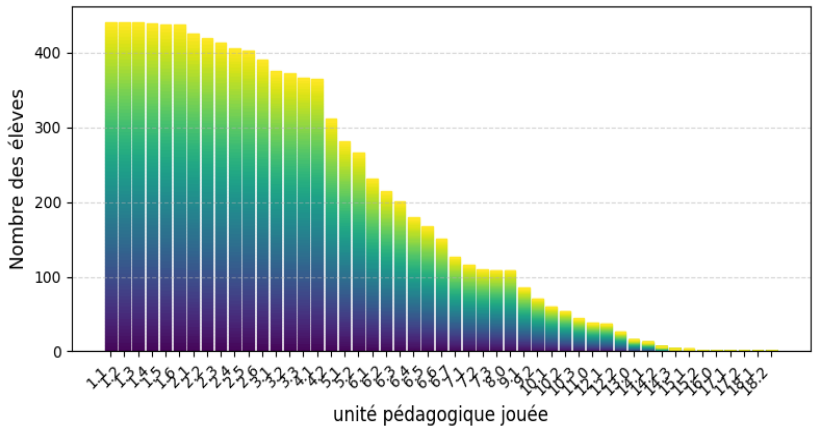



FIGURE 1 – Nombre d'élèves ayant joué à chaque unité pédagogique parmi les 440 joueurs retenus.

Une série d'évaluations a été conduite, avant et après l'entraînement des élèves avec GGF, pour mesurer l'évolution des performances en lecture au fil du temps. Les pré-tests ont été effectués en

novembre 2017, suivis d'une session post-test en juin 2018, durant laquelle les mêmes exercices ont été reproduits afin d'évaluer les progrès réalisés. Les tests d'évaluation se sont principalement axés sur diverses compétences en lecture, telles que détaillées dans (Lassault *et al.*, 2022).

Les évaluations pré- et post-intervention reposent principalement sur trois mesures : le nombre de mots correctement lus dans un texte dénué de sens lors du test de l'Alouette, les performances en décodage de pseudo-mots et de mots familiers issues du test "Lecture une Minute" LUM, ainsi que la capacité d'identification orthographique de mots, évaluée par le test TIME2.

Trois groupes d'élèves ont été identifiés sur la base de leurs performances initiales (lors des pré-tests) et le gain obtenu (post-pre) à la fin de l'année scolaire : 'les bons répondants' (cluster $C_1$), ayant enregistré des progrès supérieurs à la moyenne, 'les mauvais répondants' (cluster $C_2$), dont les performances sont demeurées significativement inférieures avec une progression limitée, et 'les bons élèves'(cluster $C_0$), qui présentaient déjà un niveau initial élevé et ont maintenu une progression continue, similaire à celle 'des bons répondants'.

L'enjeu applicatif principal de cet article[4] consiste alors à retrouver les trois groupes préalablement définis, en s'appuyant sur les indicateurs comportementaux du jeu plutôt que sur les résultats aux tests de lecture. L'identification rapide des élèves qui répondent ou pas à une intervention, telle que GGF, est capitale sur le plan pédagogique ou thérapeutique (Ziegler *et al.*, 2020).

# 4 Méthodologie à base de réseaux neuronaux graphiques

Étant donné que les joueurs ont tendance à suivre des parcours d'apprentissage non linéaires, divers scénarios d'interaction peuvent être détectés et analysés. Ces données non structurées génèrent, pour chaque joueur, une séquence d'horodatages (*timestamps*) de taille variée. Les réseaux de neurones graphiques (GNNs) sont de bons candidats pour le traitement de telles structures de données du fait de leurs représentations sous forme de graphes. Les approches de modélisation associées ne reposent pas uniquement sur la topologie, mais intègrent également les attributs des nœuds.

## 4.1 Approche GraphSAGE

GraphSAGE (*Graph Sample and Aggregation*) constitue un modèle performant pour l'apprentissage sur les graphes, basé sur une agrégation récursive des informations issues des nœuds voisins (Hamilton *et al.*, 2017). Cette approche permet de générer des représentations vectorielles (*embeddings*) pour des nœuds jamais observés durant l'apprentissage, et par conséquent pour des graphes de traces de joueurs entièrement nouveaux.

L'application d'une approche graphique pour la classification supervisée nécessite le recours à une bibliothèque d'implémentation adaptée. Spektral[5] en fournit une version robuste, compatible avec les modèles convolutionnels inductifs. Son emploi implique, toutefois, un prétraitement des données, des matrices de caractéristiques des nœuds et des structures d'adjacence. Pour cela, les niveaux joués sont modélisés comme des nœuds du graphe, tandis que les arêtes traduisent les transitions entre ces niveaux. La figure 2 illustre un exemple de parcours d'un joueur, représenté par un sous-ensemble de

4. cette étude s'inscrit dans une démarche de science ouverte.
5. https://graphneural.network

niveaux joués, dont certains sont revisités à plusieurs reprises. Ces répétitions sont symbolisées par des boucles de nœuds, accentuées par des liens en gras, soulignant ainsi les aspects non linéaires du parcours.

À partir de cette modélisation, nous avons implémenté un modèle GraphSAGE dont l'architecture repose sur trois couches GraphSageConv suivies d'un GlobalAvgPool, permettant l'apprentissage de représentations vectorielles des graphes à partir des voisinages locaux des nœuds. Les deux premières couches convolutionnelles utilisent des embeddings intermédiaires de dimension 64, tandis que la dernière projette les représentations vers un espace aligné avec le nombre de classes cibles. Une couche dense softmax assure ensuite la prédiction des profils comportementaux à partir de ces représentations.

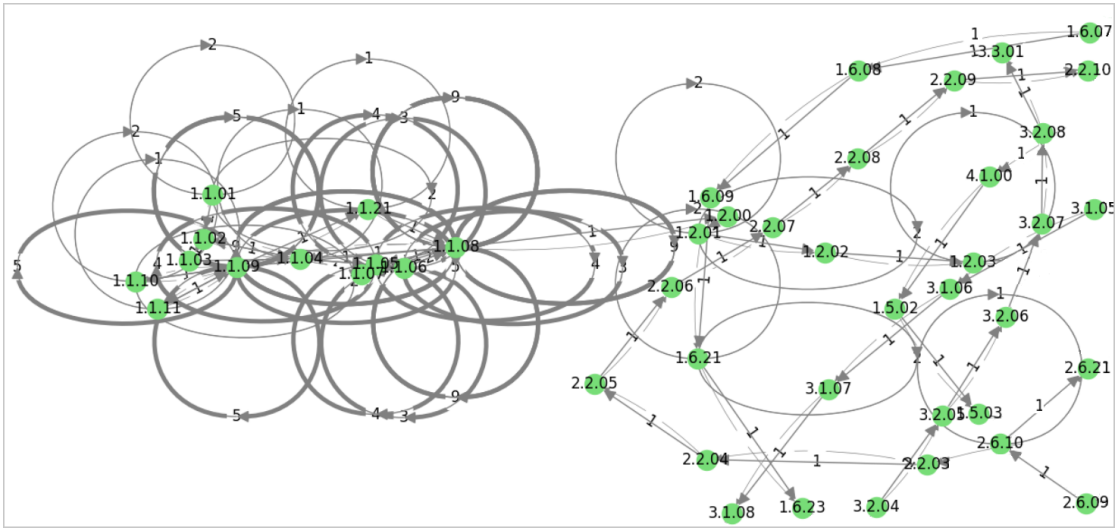



FIGURE 2 – Exemple des interactions d'un joueur pour certains niveaux joués.

En entrée, l'architecture mobilise une matrice d'adjacence creuse, encodant uniquement les connexions effectives entre les entités, une matrice d'attributs de nœuds dérivée de métriques clés du *gameplay*, décrivant les interactions par niveau, ainsi que des labels de classes définis. Malheureusement, la couche GraphSageConv n'intègre pas les poids des arêtes et repose sur une matrice d'adjacence binaire, indiquant si un lien existe ou non entre les nœuds. Cette limitation réduit la capacité du modèle à refléter des informations structurelles fines, notamment lorsque certains niveaux font l'objet de multiples retours.

## 4.2 Approche ECCCConv

En réponse à cette contrainte, une seconde architecture a été développée reposant sur la couche ECCConv (*Edge-Conditioned Convolution*) (Simonovsky & Komodakis, 2017) de Spektral, permettant d'intégrer la fréquence des transitions entre niveaux en tant que caractéristiques portées par les arêtes. Ces attributs sont ensuite exploités par cette couche pour adapter dynamiquement le filtrage appliqué aux nœuds voisins, offrant ainsi une modélisation plus fine et contextuelle des relations au sein du graphe.

L'architecture adoptée reprend la structure de GraphSAGE, avec trois couches ECCConv successives, suivies d'une opération de *pooling* global moyen. Le modèle ECCConv adopte une structure hiérarchique avec des dimensions croissantes d'embeddings successives (32, 64, puis 128), reflétant l'approche classique d'augmentation progressive de la capacité représentative. En sortie, deux couches denses sont utilisées : la première (ReLU) pour l'apprentissage de représentations profondes et la seconde (Softmax) pour la classification multi-classe. L'entrée du modèle est similaire à l'approche précédente, enrichie des attributs d'arêtes.

# 5 Résultats

Pour la tâche de classification, consistant à reconnaître à quelle classe appartient un élève parmi $C_0$, $C_1$ et $C_2$ (voir section 3), la précision obtenue est de 66% pour GraphSAGE et de 68% pour ECCConv. En complément, le F1-score met en évidence la meilleure capacité du second à distinguer les classes minoritaires, tout en maintenant de bonnes performances globales. Cette tendance en faveur d'ECCConv est également observée lors de plusieurs exécutions indépendantes (moyennes de 67,2% vs 64,9%), avec une $p - value$ de 0,089 (test $t$ de $Welch$) suggérant qu'elle mérite d'être confirmée par des analyses complémentaires.

Une stratégie de fusion pondérée (50-50) des prédictions des deux modèles a ensuite été appliquée. Cette combinaison a permis d'accroître la robustesse de la classification, atteignant une précision de 71% et améliorant la détection des profils de joueurs, ce qui renforce l'hypothèse de l'impact éducatif positif de GGF.

Une analyse par classe (cf. table 2) montre que l'approche combinée capte plus finement les spécificités de chaque profil, notamment pour les élèves en difficulté (classe $C_2$) suggérant un besoin d'accompagnement ciblé et ouvrant la voie à des parcours d'apprentissage plus adaptés. L'identification de la classe $C_1$ (F1-score = 0,68) confirme que ces élèves ont bénéficié de l'entraînement avec GGF, bien que leur distinction avec la classe $C_0$ demeure plus subtile, en raison de comportements d'apprentissage similaires.

| | **GraphSAGE** | **ECCConv** | **Approche combinée** | | |
|---|---|---|---|---|---|
| Accuracy | 66,25 | 68,75 | 71,25 | 0,65 | $C_0$ |
| | | | | 0,71 | $C_1$ |
| | | | | 0,77 | $C_2$ |
| F1-score | 64,58 | 68,74 | 71,20 | 0,70 | $C_0$ |
| | | | | 0,68 | $C_1$ |
| | | | | 0,77 | $C_2$ |

TABLE 2 – Résultats des deux approches GraphSAGE et ECCConv.

Les élèves de la classe $C_0$, ayant entamé l'expérimentation avec un bon niveau initial, obtiennent un F1-score de 0,70 indiquant une reconnaissance fiable. Ce résultat souligne que ces profils, bien qu'en minorité, sont correctement identifiés, renforçant l'utilité pédagogique du jeu pour tous les types d'élèves.

GGF permet également une analyse fine de la progression des élèves tout au long de l'expérimentation, comme l'illustre la figure 3. Celle-ci met en évidence l'évolution des traces comportementales des joueurs, segmentés en clusters, au fil du temps et selon divers indicateurs de *gameplay*.

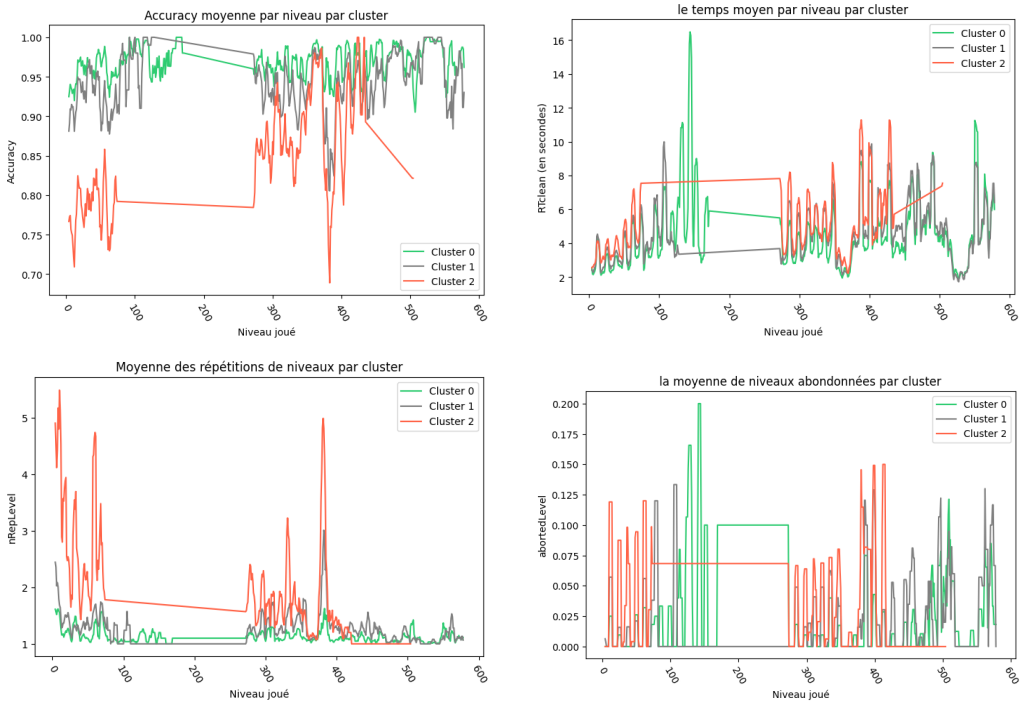



FIGURE 3 – Analyse des parcours de jeu associés aux trois clusters.

Les courbes révèlent une forte similarité de progression entre les clusters $C_0$ et $C_1$ , en contraste avec les élèves en difficulté, marqués par un faible taux de bonnes réponses et des répétitions fréquentes de niveaux sans progrès notable. À l'inverse, certains niveaux, impliquant des temps de réponse plus élevés ou des abandons en cours de tâche chez les bons élèves, restent inaccessibles aux élèves en difficulté et partiellement explorés par les bons répondants. Cela suggère que ces niveaux, probablement plus exigeants sur le plan pédagogique, constituent des points de blocage pour les élèves, et justifieraient une attention particulière en classe ou une intégration plus progressive au sein du dispositif d'apprentissage.

# 6 Conclusion et perspectives

Cette étude visait à évaluer l'impact éducatif de GraphoGameFrançais (GGF) à travers une analyse comportementale fondée sur des représentations graphiques des parcours d'apprenants. En mobilisant deux modèles de réseaux de neurones graphiques, GraphSAGE et ECCConv, nous avons pu classifier les profils d'élèves à partir des dynamiques d'apprentissage extraites, mettant ainsi en évidence les bénéfices spécifiques de l'application dans le développement des compétences en lecture.

Au-delà de sa fonction d'entraînement à la lecture, GGF se révèle également être un dispositif de suivi pédagogique efficace, permettant d'identifier les élèves nécessitant un accompagnement éducatif particulier, ainsi que ceux qui tirent pleinement parti de l'entraînement. Ce potentiel de diagnostic

ouvre la voie vers une stratégie d'apprentissage différencié, durant laquelle des interventions et des parcours adaptés pourraient être construits selon les besoins spécifiques des élèves.

Bien que notre étude ait mis en évidence l'effet éducatif de GGF, son usage en parallèle des enseignements scolaires suggère qu'il pourrait constituer un outil complémentaire aux apprentissages en classe, plutôt qu'un vecteur d'acquisition autonome. Pour mieux isoler son impact propre, il serait pertinent de mener des expérimentations dans un cadre non scolaire, auprès de groupes d'élèves aux profils homogènes (notamment en termes d'environnement familial, ou de capacités cognitives) afin de valider de manière plus contrôlée l'efficacité du dispositif.

Une dimension supplémentaire concerne la modélisation des dynamiques d'apprentissage. Elle pourrait être affinée en remplaçant les nœuds actuels (niveaux joués) par des sous-niveaux, tels que les réponses cibles "*targets*" (réponses correctes), afin de mieux exploiter les données. Toutefois, ces éléments se répètent dans différents niveaux, ce qui pourrait diluer l'aspect chronologique. Une alternative consisterait à les concaténer avec d'autres variables contextuelles, comme les options affichées à l'écran, créant ainsi des nœuds de graphes plus détaillés. Cela offrirait une représentation plus fine des parcours et permettrait de mieux capter la progression et les comportements des joueurs.

Un autre enjeu majeur porte sur la présence de variables binaires pour certains attributs. En adoptant un recodage approprié et en réservant une valeur spécifique au padding, il deviendrait possible d'exploiter des architectures séquentielles telles que les LSTM, offrant ainsi une modélisation plus affinée de la dimension temporelle.

# 7 Financement

Cette étude a été financée par le programme eFRAN (France 2030) soutenu par l'Agence Nationale de la Recherche (ANR-22-FRAN-0004), l'Institut de Convergences sur le Langage, la Communication et le Cerveau (ILCB, ANR-16-CONV-0002) et le Pôle pilote pour la recherche en éducation et la formation des enseignants (AMPIRIC).

# Références

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DRACHEN A., CANOSSA A. & SØRENSEN J. R. M. (2013). Gameplay metrics in game user research : Examples from the trenches. In *Game analytics : Maximizing the value of player data*, p. 285–319. Springer.

GHARBI H., ELAACHAK L. & FENNAN A. (2024). Replicating video game players'behavior through deep reinforcement learning algorithms. *Journal of Theoretical and Applied Information Technology*, **102**(15).

HAMILTON W., YING Z. & LESKOVEC J. (2017). Inductive representation learning on large graphs. *Advances in neural information processing systems*, **30**.

HOOSHYAR D., YOUSEFI M. & LIM H. (2018). Data-driven approaches to game player modeling : a systematic literature review. *ACM Computing Surveys (CSUR)*, **50**(6), 1–19.

JAGIRDAR H., TALWADKER R., PAREEK A., AGRAWAL P. & MUKHERJEE T. (2024). Explainable and interpretable forecasts on non-smooth multivariate time series for responsible gameplay. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 5126–5137.

KIM Y. B., KANG S. J., LEE S. H., JUNG J. Y., KAM H. R., LEE J., KIM Y. S., LEE J. & KIM C. H. (2015). Efficiently detecting outlying behavior in video-game players. *PeerJ*, **3**, e1502.

LAAMARTI F., EID M. & EL SADDIK A. (2014). An overview of serious games. *International Journal of Computer Games Technology*, **2014**(1), 358152.

LASSAULT J., SPRENGER-CHAROLLES L., ALBRAND J.-P., ALAVOINE E., RICHARDSON U., LYYTINEN H. & ZIEGLER J. C. (2022). Testing the effects of graphogame against a computer-assisted math intervention in primary school. *Scientific Studies of Reading*, **26**(6), 449–468.

LEE K. M. & PENG W. (2006). What do we know about social and psychological effects of computer games ? a comprehensive review of the current literature. *Playing video games : motives, responses, and consequences*.

MELO S. A., KOHWALTER T. C., CLUA E., PAES A. & MURTA L. (2020). Player behavior profiling through provenance graphs and representation learning. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, p. 1–11.

MENÉNDEZ H. D., VINDEL R. & CAMACHO D. (2014). Combining time series and clustering to extract gamer profile evolution. In *Computational Collective Intelligence. Technologies and Applications : 6th International Conference, ICCCI 2014, Seoul, Korea, September 24-26, 2014. Proceedings 6*, p. 262–271 : Springer.

RUIZ J.-P., LASSAULT J., SPRENGER-CHAROLLES L., RICHARDSON U., LYYTINEN H. & ZIEGLER J. C. (2017). Graphogame : un outil numérique pour enfants en difficultés d'apprentissage de la lecture. *ANAE Approche neuropsychologique des apprentissages chez l'enfant*.

SAAS A., GUITART A. & PERIÁNEZ A. (2016). Discovering playing patterns : Time series clustering of free-to-play game data. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, p. 1–8 : IEEE.

SAPIENZA A., GOYAL P. & FERRARA E. (2019). Deep neural networks for optimal team composition. *Frontiers in big Data*, **2**, 14.

SHAH J. & THUE D. (2023). Representing player behaviour via graph embedding techniques : A case study in dota 2. In *2023 IEEE Conference on Games (CoG)*, p. 1–8 : IEEE.

SHAHZAD FAROOQ S., FIAZ M., MEHMOOD I., KASHIF BASHIR A., NAWAZ R., KIM K. & KI JUNG S. (2021). Multi-modal data analysis based game player experience modeling using lstm-dnn. *Computers, Materials and Continua*, **68**(3), 4087–4108.

SIFA R., BAUCKHAGE C. & DRACHEN A. (2014). The playtime principle : Large-scale cross-games interest modeling. In *2014 IEEE conference on computational intelligence and games*, p. 1–8 : IEEE.

SIMONOVSKY M. & KOMODAKIS N. (2017). Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 3693–3702.

SMERDOV A., SOMOV A., BURNAEV E. & STEPANOV A. (2023). Ai-enabled prediction of video game player performance using the data from heterogeneous sensors. *Multimedia Tools and Applications*, **82**(7), 11021–11046.

SMITH P. (2008). Serious games 101. *Rta. Nato. Int*, p. 1–12.

WANG T., HONARI-JAHROMI M., KATSAROU S., MIKHEEVA O., PANAGIOTAKOPOULOS T., ASADI S. & SMIRNOV O. (2024). player2vec : A language modeling approach to understand player behavior in games. *arXiv preprint arXiv :2404.04234*.

WONG W. L., SHEN C., NOCERA L., CARRIAZO E., TANG F., BUGGA S., NARAYANAN H., WANG H. & RITTERFELD U. (2007). Serious video game effectiveness. In *Proceedings of the international conference on Advances in computer entertainment technology*, p. 49–55.

ZIEGLER J. C., PERRY C. & ZORZI M. (2020). Learning to read and dyslexia : From theory to intervention through personalized computational models. *Current Directions in Psychological Science*, **29**(3), 293–300.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Recommandation de tests multi-objectifs pour l'apprentissage adaptatif

Nassim Bouarour[1] Idir Benouaret[2] Sihem Amer-Yahia[1]
(1) LIG, 700 Av. Centrale, 38401 Saint-Martin-d'Hères, France
(2) LRE, 14-16 Rue Voltaire, 94270 Le Kremlin-Bicêtre, France
nassim.bouarour@univ-grenoble-alpes.fr, idir.benouaret@epita.fr,
sihem.amer-yahia@univ-grenoble-alpes.fr

## RÉSUMÉ

L'amélioration des compétences (*upskilling*) est un segment en forte croissance en éducation. Pourtant, peu de travaux algorithmiques se concentrent sur l'élaboration de stratégies dédiées pour atteindre une maîtrise avancée des compétences. Dans cet article, nous formalisons AdUp, un problème d'amélioration itérative des compétences combinant l'apprentissage par maîtrise et la théorie de la Zone de Développement Proximal. Nous étendons nos travaux précédents et concevons deux solutions pour AdUp : `MOO` et `MAB.MOO` est une approche d'optimisation multi-objectifs qui utilise une méthode de *Hill Climbing* pour adapter la difficulté des tests recommandés selon 3 objectifs : la performance prédite de l'apprenant, son aptitude, et son gap. `MAB` est une approche basée sur les bandits manchots (*Multi-Armed Bandits*) permettant d'apprendre la meilleure combinaison d'objectifs à optimiser à chaque itération. Nous montrons comment ces solutions peuvent être couplées avec deux modèles courants de simulation d'apprenants : BKT et IRT. Nos expérimentations démontrent la nécessité de prendre en compte les 3 objectifs et d'adapter dynamiquement les objectifs d'optimisation aux capacités de progression de l'apprenant, car `MAB` permet un taux de maîtrise plus élevé.

## ABSTRACT

**Multi-objective Test Recommendation for Adaptive Learning**

Upskilling is a fast-growing segment of the education economy. Yet, there is little algorithmic work that focuses on crafting dedicated strategies to reach high-skill mastery. In this paper, we formalize ADUP, an iterative upskilling problem that combines mastery learning and Zone of Proximal Development. We design two solutions for ADUP : `MOO` and `MAB`. `MOO` is a multi-objective optimization approach that relies on Hill Climbing to adapt the difficulty of recommended tests to three objectives : learner's predicted performance, aptitude, and skill gap. `MAB` is a meta approach based on Multi-Armed Bandits to learn the best combination of objectives to optimize at each iteration. We show how these solutions are combined with two common learner simulation models : BKT (KT-IDEM) and Item Response Theory (IRT). Our simulation experiments demonstrate the necessity of leveraging all three objectives and the need to adapt the optimization objectives to the learner's progression ability as `MAB` offers a higher mastery rate and a better final skill gain than `MOO`.

MOTS-CLÉS : Apprentissage adaptatif, recommandation, optimisation..

KEYWORDS: Adaptive learning, recommmendation, optimization..

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 94–97.

# 1 Contexte

La croissance rapide des nouvelles opportunités d'apprentissage comme les MOOCs, les tutoriels ou les forums communautaires oriente de plus en plus l'attention vers l'amélioration des compétences en ligne. L'*upskilling* (montée en compétences) en dehors des parcours de formation formels constitue aujourd'hui un segment en forte expansion de l'économie de l'éducation. Par ailleurs, les apprenants sont de plus en plus engagés dans un apprentissage auto-dirigé, gérant eux-mêmes de nombreux aspects de leur formation, ce qui implique souvent de travailler de manière autonome sur diverses activités d'apprentissage. Par conséquent, il devient de plus en plus difficile de garantir la qualité des acquis dans ces formats d'apprentissage, car ils peuvent entraîner une compréhension superficielle d'un sujet.

Idéalement, chaque apprenant devrait recevoir des tests choisis de façon à faire progresser réellement ses compétences, en tenant compte de sa capacité à résoudre des exercices en fonction de son niveau et de ses performances passées. C'est précisément l'objectif de l'apprentissage par maîtrise, une stratégie pédagogique qui met l'accent sur le temps nécessaire à chaque apprenant pour acquérir les mêmes compétences et atteindre le même niveau de maîtrise.

# 2 Contributions

Ce travail publié à *Trans. Large Scale Data Knowl. Centered Syst'24* (Bouarour *et al.*, 2024) prolonge un travail antérieur (Bouarour *et al.*, 2023) en formalisant AdUp (*Adaptive Upskilling*) comme un problème d'optimisation où l'on cherche, à chaque itération, à recommander un ensemble de k-tests à un apprenant. Ces tests doivent maximiser la performance attendue et l'aptitude, tout en minimisant l'écart de compétence accumulé. La combinaison simultanée de ces trois objectifs constitue la nouveauté principale de cette formalisation. Le défi majeur réside dans la nature multi-objectifs du problème. Deux solutions sont explorées :

— `MOO` (*Multi-Objective Optimization*) : Cette approche repose sur une solution de Pareto basée sur la dominance entre ensembles de tests, et utilise une heuristique de type *Hill Climbing* (Omidvar-Tehrani *et al.*, 2016) pour trouver des solutions non dominées. Diverses variantes peuvent être construites selon les combinaisons d'objectifs. Toutefois, MOO applique les mêmes objectifs tout au long du processus, ce qui limite son adaptabilité.

— `MAB` (*Multi-Armed Bandits*) : Pour pallier cette rigidité, `MAB` est introduite comme une approche adaptative qui apprend dynamiquement quels objectifs optimiser à chaque itération. Par exemple, si un apprenant échoue plusieurs fois aux mêmes tests, il serait plus pertinent de privilégier la réduction de l'écart avant de proposer des tests plus difficiles. `MAB` est ainsi formalisée comme un problème de type bandit manchot, capable d'adapter sa stratégie au comportement de l'apprenant.

# 3 Expérimentations

Nos expériences visent à évaluer l'efficacité des dimensions d'optimisation sur la montée en compétence (upskilling). Pour cela, nous avons divisé notre étude expérimentale en deux parties. Dans la

première partie, nous analysons l'impact de nos solutions sur l'atteinte de la maîtrise en simulant les réponses des apprenants ainsi que l'ensemble du processus d'apprentissage. Nous formulons quatre questions de recherche :

- **RQ1.** La combinaison de toutes les dimensions d'optimisation est-elle bien adaptée pour atteindre la maîtrise et améliorer la progression des compétences ?
- **RQ2.** Les paramètres choisis pour la stratégie de mise à jour des compétences influencent-ils les résultats ?
- **RQ3.** Le choix du modèle de simulation de l'apprenant : Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1994) et Item Response Theory (IRT) (Reckase & Reckase, 2009), a-t-il un impact sur la maîtrise et la progression ?
- **RQ4.** L'application d'une méta-stratégie (comme `MAB`), qui choisit dynamiquement un sous-ensemble de dimensions à optimiser à chaque itération, améliore-t-elle l'atteinte de la maîtrise ?

Les résultats expérimentaux montrent que l'approche `MOO` permet d'atteindre le plus haut taux de maîtrise en un nombre réduit d'itérations, confirmant ainsi la théorie de la Zone de Développement Proximal (ZPD) et du Flow, ainsi que l'importance d'exploiter l'aptitude pour proposer des défis adaptés aux apprenants. Notre étude révèle que `MOO` est robuste face aux variations du paramètre N dans la stratégie de mise à jour des compétences (N réponses correctes consécutives). Les conclusions obtenues avec le modèle de simulation BKT sont également généralisables avec IRT, bien que ce dernier favorise la réduction de l'écart, tandis que BKT privilégie la performance attendue.

# 4 Conclusion

Nous avons abordé la montée en compétence adaptative en suivant une approche d'apprentissage par maîtrise. L'originalité de notre méthode réside dans l'adaptation de la difficulté des tests selon les performances prédites de l'apprenant, son aptitude et ses lacunes. Deux approches ont été proposées : `MOO`, qui résout directement le problème, et `MAB`, qui choisit dynamiquement entre différentes variantes d'optimisation à chaque itération. Nos expériences ont montré que `MAB` permet un meilleur taux de maîtrise et une progression plus importante des compétences finales par rapport à `MOO`. Cependant, `MOO` attribue des tests de meilleure qualité et avec plus de précision. Nous avons également testé l'effet de différents modèles de simulation d'apprenants sur la réussite à la maîtrise.

Dans le futur, nous envisageons d'enrichir notre cadre théorique en intégrant d'autres théories de l'apprentissage, telle que l'apprentissage collaboratif, qui a prouvé son efficacité en ligne, notamment via le feedback entre pairs et les discussions entre apprenants. Enfin, nous visons une personnalisation plus fine de l'expérience d'apprentissage en modélisant des profils d'apprenants à partir de leurs performances passées. Ces profils pourraient être utilisés pour attribuer des tests soit par regroupement selon leur niveau général, soit par filtrage collaboratif basé sur les réussites d'apprenants similaires.

# Références

BOUAROUR N., BENOUARET I. & AMER-YAHIA S. (2024). *Multi-objective Test Recommendation for Adaptive Learning*, In A. HAMEURLAIN, A. M. TJOA, R. AKBARINIA & A. BONIFATI, Éds.,

*Transactions on Large-Scale Data- and Knowledge-Centered Systems LVI : Special Issue on Data Management - Principles, Technologies, and Applications*, p. 1–36. Springer Berlin Heidelberg : Berlin, Heidelberg. DOI : 10.1007/978-3-662-69603-3_1.

BOUAROUR N., BENOUARET I., D'HAM C. & AMER-YAHIA S. (2023). Adaptive Test Recommendation for Mastery Learning. In *Proceedings of the 2nd International Workshop on Data Systems Education : Bridging Education Practice with Education Research*, DataEd '23, p. 18–23, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3596673.3596977.

CORBETT A. T. & ANDERSON J. R. (1994). Knowledge tracing : Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, **4**, 253–278.

OMIDVAR-TEHRANI B., AMER-YAHIA S., DUTOT P.-F. & TRYSTRAM D. (2016). Multi-objective group discovery on the social web. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, p. 296–312 : Springer.

RECKASE M. D. & RECKASE M. D. (2009). Unidimensional item response theory models. *Multidimensional item response theory*, p. 11–55.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Repenser les pratiques d'enseignement et d'apprentissage par la robotique éducative : le cas du robot socio-émotionnel Buddy

Ismail Badache[1] Elisabeth Colombo[2]
(1) Aix-Marseille University, Université de Toulon, CNRS, LIS, Marseille, France
(2) L'odyssée d'Elise, 37, rue Chardon Lagache 75016, Paris, France
ismail.badache@univ-amu.fr, contact@elisabeth-colombo.fr

## RÉSUMÉ

Cet article explore l'utilisation de **Buddy** dans un contexte éducatif et d'apprentissage, avec un focus particulier sur deux usages. Premièrement, à l'Institut National Supérieur du Professorat et de l'Éducation d'Aix-Marseille avec des étudiants futurs profs des écoles, collèges et lycées ainsi que des étudiants en ingénierie pédagogique numérique. Deuxièmement, dans un contexte spécifique comme médiateur artistique et émotionnel dans l'apprentissage de l'art. Cet article s'intéresse à la façon dont ce robot peut enrichir les pratiques pédagogiques en stimulant la créativité, l'interaction et l'accompagnement pédagogique et émotionnel des apprenants. **Buddy** peut agir comme médiateur entre l'apprenant et son environnement, en particulier dans les domaines de la narration, de l'art et de l'assistance informationnelle. Cette expérimentation du robot **Buddy**, met en lumière les possibilités de la robotique dans le développement de pratiques pédagogiques inclusives, où l'art et la technologie convergent pour favoriser l'apprentissage et la résilience émotionnelle. À travers ces expériences, le robot devient un catalyseur d'apprentissage et de réflexion, tout en ouvrant des perspectives pour une recherche interdisciplinaire impliquant l'ingénierie informatique, la psychologie et l'éducation. Les limites techniques actuelles de **Buddy**, loin d'être des obstacles, offrent des opportunités pour concevoir des scénarios pédagogiques visant à démythifier l'intelligence artificielle, en mettant en lumière ses biais et ses limitations.

## ABSTRACT

**Buddy : Rethinking Teaching and Learning Practices through Educational Robotics**

This paper explores the use of **Buddy** in educational and learning contexts, focusing on two key uses. First, at the National Higher Institute of Teaching and Education in Aix-Marseille, with future teachers and digital pedagogy students. Second, as an emotional and artistic mediator in art learning. The paper examines how **Buddy** can enhance teaching practices by stimulating creativity, interaction, and emotional and educational support. Acting as a mediator between the learner and their environment, particularly in storytelling, art, and informational assistance, **Buddy** highlights the potential of robotics to develop inclusive pedagogical practices where art and technology converge to promote learning and emotional resilience. Through these experiments, **Buddy** serves as a learning catalyst, opening avenues for interdisciplinary research involving computer engineering, psychology, and education. The current technical limitations of **Buddy** offer opportunities to design educational scenarios that demystify artificial intelligence, shedding light on its biases and limitations.

MOTS-CLÉS : Robotique éducative et émotionnelle, Éducation artistique, IA, Pédagogie.
KEYWORDS: Educational and Emotional Robotics, Art education, AI, Pedagogy.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 98–110.

# 1 Introduction

Dans un contexte éducatif en pleine mutation, où l'expression émotionnelle et la créativité sont de plus en plus reconnues comme des compétences essentielles à développer chez les élèves (UNESCO, 2021), les approches traditionnelles peinent parfois à favoriser l'engagement, l'estime de soi et la résilience émotionnelle. À la croisée des arts, de la technologie et des sciences de l'éducation, la robotique émotionnelle émerge comme une piste prometteuse pour réconcilier innovation technopédagogique et humanisme éducatif.

Les robots sociaux comme **Buddy**, conçus pour interagir de manière affective et empathique, permettent d'envisager de nouvelles modalités d'enseignement, particulièrement adaptées aux enjeux de l'éducation artistique et socio-émotionnelle. En s'appuyant sur la capacité de ces artefacts technologiques à médiatiser l'expression personnelle et à stimuler des processus réflexifs, la robotique émotionnelle devient un levier pour renforcer l'autonomie affective, la communication interpersonnelle et l'intelligence émotionnelle des élèves.

Ce article s'inscrit dans une dynamique de recherche centrée sur l'intégration des technologies robotiques dans les dispositifs pédagogiques, en particulier dans le cadre de l'interaction homme-robot (IHR) appliquée à l'éducation. Il vise à analyser les usages pédagogiques du robot social **Buddy** à travers deux études préliminiaires exploratoires distinctes mais complémentaires. L'approche adoptée s'inscrit dans une perspective interdisciplinaire mobilisant à la fois les sciences de l'éducation et l'informatique.

- **L'expérience 1**, centrée sur l'éducation artistique, mobilise **Buddy** comme compagnon de création et facilitateur de narration émotionnelle. À travers des activités artistiques intégrant l'expression corporelle, la voix et le dessin, les élèves sont invités à projeter leurs émotions et à construire une histoire avec le robot, en s'appuyant sur des mécanismes de reconnaissance affective et de feedback empathique.
- **L'expérience 2** prolonge cette démarche en intégrant **Buddy** dans des situations pédagogiques visant le développement de compétences socio-émotionnelles plus larges, notamment l'identification des émotions, la régulation affective et l'empathie, à travers des scénarios narratifs interactifs et des échanges simulés.

Ces deux expériences, bien que distinctes dans leur ancrage disciplinaire (artistique d'une part, socio-émotionnel et éducatif de l'autre), sont articulées autour d'une même hypothèse : le potentiel du robot **Buddy** à agir comme médiateur éducatif favorisant un environnement d'apprentissage bienveillant, expressif et propice au développement des élèves.

L'article s'organise en quatre temps complémentaires : il débute par une revue de littérature portant sur les apports de la robotique sociale et émotionnelle en éducation, en s'appuyant sur les travaux récents en sciences de l'éducation et en interaction homme-machine ; il présente ensuite en détail les deux expériences pédagogiques menées avec le robot **Buddy**, en décrivant leurs contextes, leurs objectifs et leurs dispositifs didactiques ; une analyse croisée des résultats permet ensuite d'identifier les effets observés sur l'engagement des élèves, la verbalisation émotionnelle et la coopération entre pairs ; enfin, une discussion met en perspective les apports, les limites et les perspectives de la robotique émotionnelle comme levier d'innovation éducative, en lien avec les enjeux contemporains d'une éducation plus inclusive, expressive et empathique.

## 2 Les robots en éducation : une vue d'ensemble

Les robots connaissent une adoption croissante dans l'utilisation des technologies multimédia, principalement en raison des avancées technologiques et de leur rôle de plus en plus central dans le domaine de l'éducation (Pachidis *et al.*, 2019). Un robot social/éducatif est un appareil autonome ou semi-autonome conçu pour interagir de manière significative avec les humains/apprenants. Grâce à l'intelligence artificielle (IA) et aux algorithmes d'apprentissage automatique dont ces robots sont dotés, ils peuvent percevoir et réagir aux émotions (simulation), actions et signaux sociaux des individus. Ces robots sont souvent programmés pour imiter des caractéristiques humaines, telles que la voix, les gestes et les expressions faciales (Breazeal, 2004b). Selon plusieurs chercheurs, certains des effets comportementaux de ce type de robots en milieu scolaire expliquent diverses formes d'interaction homme-robot. Ainsi, des robots avec des fonctionnalités spécifiques pourraient être employés comme outils pédagogiques pour initier les élèves aux technologies (Tang & Chu, 2022).

Les robots sociaux, à la différence des robots éducatifs classiques, possèdent la capacité d'interpréter et d'exprimer des émotions simulées, ce qui leur confère un potentiel transformateur dans les contextes éducatifs. Leur introduction en milieu scolaire est susceptible de modifier les pratiques pédagogiques, de favoriser des approches d'enseignement personnalisées, d'influer sur les résultats d'apprentissage et de redéfinir le rôle de l'enseignant. En tant qu'entités sociales, ils reposent sur les avancées de l'IA et de l'interaction homme–machine, intégrant une interface sociale qui, conjuguée à leur apparence anthropomorphe et à leur fonction relationnelle, les rend perçus comme des agents sociaux à part entière (Hegel *et al.*, 2009).

Dotés d'une aptitude à interagir socialement, ces robots peuvent établir des communications, apprendre et s'adapter à leurs interlocuteurs humains ou artificiels (Breazeal, 2004a; Fong *et al.*, 2003). Pour favoriser leur acceptation et leur déploiement à plus grande échelle, il est essentiel de comprendre ces dynamiques interactionnelles (Hegel *et al.*, 2009). Leur conception anthropomorphique met en avant leur intelligence émotionnelle, leurs compétences socio-cognitives, leur capacité d'incarnation physique et leur orientation vers des échanges interpersonnels riches et significatifs (Breazeal *et al.*, 2016).

En tant qu'agents affectifs, les robots sociaux sont capables de détecter et de manifester des variations émotionnelles, affichant des compétences sociales avancées, une réceptivité aux signaux sociaux et une sociabilité marquée (Kirby *et al.*, 2010; Breazeal, 2003). De ce fait, l'analyse de la qualité de leur engagement avec les usagers devient une nécessité pour mesurer leur impact réel (Anzalone *et al.*, 2015).

Ces caractéristiques font des robots sociaux des candidats sérieux pour une intégration efficace non seulement dans les systèmes éducatifs, mais également dans d'autres domaines d'application (Leite *et al.*, 2013). Cependant, face à la multiplicité des modèles existants (Nao, **Buddy**, Pepper, Haru, Qtrobot, Kebbi, etc.), aux fonctionnalités diverses, le choix du robot le plus pertinent pour chaque contexte d'apprentissage doit être effectué avec rigueur (Mahdi *et al.*, 2022).

Contrairement aux technologies éducatives dématérialisées, les robots sociaux se distinguent par leur présence physique, laquelle joue un rôle déterminant dans l'augmentation des performances scolaires, notamment sur les plans affectif et cognitif, en particulier lorsqu'ils endossent le rôle de tuteurs ou de pairs d'apprentissage (Belpaeme *et al.*, 2018). Lorsqu'ils sont assignés à des tâches ciblées, leurs effets pédagogiques peuvent se révéler comparables à ceux observés avec des enseignants humains (Belpaeme *et al.*, 2018; Woo *et al.*, 2021).

Dans cette perspective, il apparaît crucial de porter une attention particulière aux comportements sociaux de ces dispositifs et à la manière dont ils interagissent avec les différents acteurs de l'éducation (van den Berghe *et al.*, 2019). Cette dimension prend une importance encore plus grande dans le cadre de l'éducation spécialisée.

Les robots sociaux peuvent ainsi devenir de véritables partenaires pédagogiques, aussi bien dans des contextes d'apprentissage formels qu'informels (Johal, 2020), avec des bénéfices identifiés tant pour les élèves que pour les enseignants (Smakman *et al.*, 2021). Ces derniers manifestent d'ailleurs des perceptions globalement favorables à leur égard.

Néanmoins, plusieurs enjeux doivent être anticipés avant une intégration généralisée de ces dispositifs dans les environnements éducatifs : enjeux éthiques et moraux (Smakman *et al.*, 2021), mais aussi défis techniques, organisationnels et économiques (Belpaeme & Tanaka, 2021). Par conséquent, il importe de tempérer les attentes parfois exagérément optimistes quant à leur potentiel disruptif dans l'enseignement. Il conviendrait plutôt d'élaborer des cadres de référence solides et des lignes directrices précises pour guider leur conception et leur intégration pédagogique (Mahdi *et al.*, 2022; Woo *et al.*, 2021), de renforcer la robustesse et la standardisation de leurs composantes matérielles et logicielles (Pachidis *et al.*, 2019), et d'analyser de manière rigoureuse les effets concrets de leur usage sur les processus d'enseignement et d'apprentissage (Barakova *et al.*, 2023).

# 3 C'est quoi Buddy ?

**Buddy** est un robot social développé par *Blue Frog Robotics*[1] (voir la figure 1), une entreprise française basée à Paris. Son objectif principal est d'être un robot de compagnie amical pour toute la famille. Il est conçu comme un nouveau type de partenaire, se situant à mi-chemin entre l'interaction humain-animal et l'interaction humain-humain. L'accent est mis sur une communication intuitive et naturelle avec l'utilisateur (Milliez, 2018).

FIGURE 1 – Vues sous différents angles du robot **Buddy**

1. https ://www.bluefrogrobotics.com/fr/

TABLE 1 – Caractéristiques générales du robot **Buddy**

| **Spécification** | **Valeur** |
|---|---|
| Dimensions (Hauteur x Largeur x Profondeur) | 560 mm x 350 mm x 350 mm |
| Poids | 8 kg |
| Batterie | Lithium-Ion < 100 Wh |

TABLE 2 – Composants et fonctionnalités du robot **Buddy**

| **N°** | **Composant/Fonctionnalité** | **Description** | **Utilité principale** |
|---|---|---|---|
| 1 | Écran tactile 8” | Interface graphique pour interaction utilisateur | Affichage, navigation, expression faciale |
| 2 | LEDs + connecteurs d’accessoires (x2) | Éclairage + ports d’extension | Signaux visuels, ajout de périphériques |
| 3 | LED cœur | LED expressive au centre du torse | Communication émotionnelle non verbale |
| 4 | Haut-parleur | Sortie audio | Voix, alertes, musique |
| 5 | Capteurs à ultrasons (x2) | Capteurs de distance à moyenne portée | Évitement d’obstacles |
| 6 | Capteurs infrarouge (x4) | Capteurs de proximité à courte portée | Navigation, sécurité |
| 7 | Caméra 13 Mpx (80°) | Caméra frontale à champ moyen | Reconnaissance faciale, interactions |
| 8 | LED d’éclairage | Éclairage frontal | Amélioration des prises de vue |
| 9 | Caméra 13 Mpx (130°) | Caméra à grand angle | Vision large, cartographie |
| 10 | Capteurs tactiles (x6) | Capteurs répartis sur le corps | Réactions au toucher |
| 11 | Microphones omnidirectionnels | Microphones captant les sons à 360° | Reconnaissance vocale, communication |
| 12 | Capteurs de vide infrarouge (x7) | Détecteurs de vide sous le robot | Prévention des chutes (escaliers, rebords) |
| 13 | Interrupteur On/Off | Bouton d’alimentation | Démarrage/arrêt du robot |
| 14 | Connecteurs d’alimentation | Port pour recharge manuelle | Recharge de la batterie |
| 15 | Connecteurs de recharge | Connexions station de recharge | Recharge automatique |
| 16 | Slot Nano SIM (4G) | Emplacement carte SIM | Connexion mobile (Internet, téléphonie) |
| 17 | Port USB | Connectique standard | Mise à jour, accessoires |
| 18 | LED état caméra | Indicateur d’activité caméra arrière | Respect de la vie privée |

Ci-dessous les caractéristiques de **Buddy** (Milliez, 2018) :

- ☑ Il mesure 56 centimètres (plus de détails sur le tableau 1) et présente un design attrayant avec une face anthropomorphe capable d’exprimer des réactions émotionnelles. Sa taille et son comportement évoquent ceux d’un animal de compagnie, ce qui renforce son accessibilité et son attrait affectif. L’ensemble de son apparence est conçu spécifiquement pour favoriser l’interaction humain-robot (HRI).

- ☑ Il est conçu pour être un compagnon pour toute la famille, en proposant un type d'interaction, situé à mi-chemin entre la relation humain-animal et humain-humain. Par ailleurs, il se veut également un outil au service de la recherche et de l'éducation, notamment grâce à un SDK dédié permettant son exploitation dans des contextes pédagogiques et scientifiques.
- ☑ Il a un comportement proactif, il peut rechercher activement un utilisateur dans son environnement et proposer des activités, ce qui renforce son rôle d'interlocuteur social et engageant.
- ☑ Il intègre des fonctions avancées de navigation et de détection de l'environnement grâce à divers capteurs (voir le tableau 2 ci-dessous), notamment un capteur Time Of Flight, des ultrasons, une caméra 3D et des capteurs de sol. Ces dispositifs lui permettent d'éviter les obstacles et les chutes, assurant ainsi des déplacements en toute sécurité dans son environnement.
- ☑ Il prend en charge une interaction multimodale avec l'utilisateur. Il est capable de traiter le langage naturel, en comprenant et en générant des énoncés pour faciliter une communication fluide. Grâce à ses micros, il peut également écouter et réagir aux commandes vocales. Enfin, il répond au toucher par l'intermédiaire de son écran tactile et de ses capteurs de caresse, permettant ainsi une interaction physique intuitive et affective.
- ☑ Il est doté d'une capacité d'expression émotionnelle qu'il manifeste à travers sa face expressive, l'utilisation de LEDs colorées, de moteurs, de sons et de sa voix de synthèse, renforçant ainsi la dimension affective de l'interaction. En parallèle, il assure la gestion d'un état interne, composé de désirs et d'émotions, qui évolue en fonction des stimuli de l'environnement et des interactions passées. Cet état influence directement ses comportements, tant sur le fond que sur la forme des actions qu'il entreprend.
- ☑ Il constitue une plateforme évolutive et collaborative, accessible via un App Store ouvert à la communauté. Il propose une variété d'applications adaptées aux besoins des différents membres de la famille. Pour les enfants, il offre des contenus d'édutainment, tels que des jeux de mémoire ou de calcul. Pour les adultes, il intègre des outils utilitaires comme un contrôleur d'objets connectés (IoT), des services de météo ou de visioconférence. De cette manière, **Buddy** s'affirme comme un compagnon polyvalent au service de tous.
- ☑ Il est livré avec un kit de développement logiciel (SDK) intuitif, basé anciennement sur le moteur de jeu Unity, qui permet aux développeurs d'accéder à l'ensemble des composants matériels et logiciels du robot. Ce SDK inclut actuellement des fonctions de haut niveau essentielles à l'interaction humain-robot (HRI), telles que la navigation autonome, l'affichage de comportements émotionnels, la gestion de l'interface graphique et la gestion des dialogues, facilitant ainsi la création d'applications personnalisées et riches en interactions.

# 4 Scénarios expérientiels avec buddy

Cette section présente deux expériences d'usage du robot **Buddy**. La première s'inscrit dans le cadre de l'éducation artistique, tandis que la seconde porte sur sa découverte et sa mise en œuvre à l'INSPE d'Aix-Marseille Université.

## 4.1 Expérience 1 : Éducation artistique

L'expérience avec **Buddy** dans le cadre de l'éducation artistique concerne deux scénarios développés pour explorer les capacités de **Buddy** à agir en tant que médiateur artistique, support de créativité et d'expression émotionnelle, ainsi qu'acteur de narration poétique. Ces 2 scénarios sont détaillés ci-dessous :

**L'émotion partagée : de l'interprétation à l'expression**
Ce scénario propose une exploration progressive de l'émotion, de la réception à la création. Dans un premier temps, des œuvres visuelles issues du travail artistique de l'auteure sont présentées. Pour chacune, **Buddy** est invité à formuler une interprétation sensible : « *Je vois du courage dans ce bleu profond...Et vous, qu'en pensez-vous ?* ». Les enfants ou participant·e·s réagissent à leur tour : à l'oral, par un mot, un dessin ou une couleur. Ce dialogue entre les émotions projetées du robot, celles ressenties par les participant·e·s, et l'intention artistique de l'auteure sert de tremplin à une deuxième phase, plus introspective : **Buddy** invite chacun à exprimer une émotion personnelle à travers une œuvre libre, en posant des questions simples et bienveillantes : « *Qu'est-ce que tu ressens aujourd'hui ? Dessine-le pour moi.* » **Buddy** accueille chaque production sans jugement, reformule parfois avec douceur (« *Tu veux dire que cette ombre te rend triste ?* »), et crée un cadre sécurisant pour l'expression de soi. Ce scénario renforce la verbalisation émotionnelle, l'écoute de l'autre, l'estime de soi et la créativité, en s'appuyant sur le rôle médiateur du robot comme **interprète sensible, miroir partiel, et facilitateur émotionnel**.

**Narration symbolique : les murmures de Buddy**
Inspiré de la série artistique de l'auteure, ce scénario met en scène **Buddy** dans un récit illustré aux côtés d'une Guerrière Écarlate, figure de résilience née de l'adversité. Chaque épisode aborde une émotion ou une lutte intérieure (solitude, doute, espoir), à travers une narration poétique et une interaction robotisée. Les élèves ou participant·e·s sont invités à créer une œuvre en lien avec le thème, tandis que **Buddy** commente, questionne ou réagit à l'émotion exprimée. Ce scénario croise storytelling, art et IA dans une pédagogie de la reconstruction et de l'introspection symbolique. Il illustre le potentiel du robot comme acteur poétique et médiateur symbolique, au service d'une pédagogie sensible et inclusive.

Les ateliers visant à dérouler les deux scénarios ont été menés dans un cadre artistique ouvert, destiné à des enfants âgés de 7 à 10 ans, ainsi que dans des contextes de médiation sensible auprès de publics inhibés. Le robot **Buddy** y était configuré pour alterner entre des comportements expressifs, des formulations empathiques et des postures symboliques inspirées du récit de la *Guerrière Écarlate*.

### 4.1.1 Quels usages artistiques ciblés ?

En mobilisant **Buddy** dans l'apprentissage, on peut renouer avec des approches pédagogiques actives et incarnées, proches de celles défendues par Seymour Papert[2] à travers le langage Logo (Papert, 1980). Là où un enfant donne un ordre abstrait à un curseur sur un écran, **Buddy** peut physiquement exécuter cet ordre, tout en le verbalisant : « *J'avance de 5 longueurs.* »

Ce type de médiation engage à la fois la mémoire verbale, la coordination motrice et la compréhension spatiale, tout en renforçant l'attention par la présence physique du robot.

Dans un contexte éducatif marqué par une préoccupation croissante pour la réduction du temps d'écran chez les enfants, **Buddy** offre une alternative crédible. Sa forme incarnée stimule les interactions orales, limite le recours aux interfaces numériques classiques, et favorise un engagement cognitif et émotionnel plus soutenu.

En cela, **Buddy** ne se contente pas d'être un outil interactif : il devient un interlocuteur pédagogique

2. https://fr.wikipedia.org/wiki/Seymour_Papert

hybride, capable de relier langage, action et perception dans une dynamique d'apprentissage ancrée dans le réel.

Ces scénarios (voir le tableau 3 ci-dessous) visent à développer des compétences artistiques, émotionnelles et réflexives. L'enjeu n'est pas seulement de produire une œuvre, mais de permettre une mise en mots ou en formes d'un ressenti, en s'appuyant sur l'intelligence émotionnelle simulée du robot.

TABLE 3 – Exemple de tableau pédagogique

| Scénario | Compétences pédagogiques mobilisées | Modalités pédagogiques |
|---|---|---|
| **L'émotion partagée : de l'interprétation à l'expression** | Reconnaissance émotionnelle, verbalisation, écoute active, créativité | Analyse d'œuvres, interaction verbale/visuelle, expression libre |
| **Narration symbolique : les murmures de Buddy** | Narration, symbolique, introspection, construction de sens par l'art | Storytelling, création artistique inspirée, dialogue poétique |

### 4.1.2 Résultats, discussion et limites

Les premiers retours issus des expérimentations avec **Buddy** mettent en évidence l'impact positif de son expressivité non verbale sur l'engagement des enfants, en particulier chez ceux présentant une réserve ou une difficulté à s'exprimer en contexte social. La capacité du robot à mobiliser des gestes expressifs, des postures corporelles cohérentes et des modulations de lumière ou de sons a contribué à instaurer un climat de confiance propice à la participation. Cette expressivité incarnée agit comme un levier d'engagement socio-émotionnel, facilitant l'identification des enfants au robot et stimulant ainsi leur implication dans les activités proposées (Belpaeme *et al.*, 2018).

Les scénarios narratifs mis en œuvre dans les séances ont par ailleurs favorisé une meilleure verbalisation émotionnelle, en offrant un cadre sécurisant et structuré pour l'expression de sentiments souvent difficiles à formuler. L'interaction avec **Buddy**, à travers des récits co-construits ou des jeux de rôle émotionnels, a permis de stimuler la créativité des enfants tout en leur offrant des opportunités d'explorer et de nommer leurs émotions, renforçant ainsi leurs compétences socio-affectives.

Comparée aux dispositifs numériques traditionnels tels que les tablettes ou les écrans, l'interaction avec un robot social doté de traits anthropomorphiques stylisés, comme **Buddy**, suscite un engagement émotionnel et cognitif distinct. En effet, le design de **Buddy**, avec son visage expressif, ses grands yeux et sa bouche animée évoquant un personnage de dessin animé, favorise une relation plus chaleureuse et accessible que celle permise par des interfaces plus abstraites ou par des robots humanoïdes réalistes, parfois perçus comme dérangeants ou flippants en raison de l'*uncanny valley* (Mori *et al.*, 2012). Ce type d'anthropomorphisme stylisé, en s'éloignant d'une imitation fidèle de l'humain, permet de susciter l'empathie sans générer de malaise, tout en facilitant la reconnaissance des émotions et l'interprétation des intentions du robot.

Cependant, ces résultats encourageants doivent être nuancés par certaines limites. L'introduction d'un robot émotionnel dans un cadre éducatif artistique soulève des questions éthiques et pédagogiques, notamment en ce qui concerne l'authenticité des émotions exprimées et le risque d'attachement excessif au robot. Certains enfants pourraient développer des formes de projection émotionnelle sur **Buddy**, attribuant à celui-ci des intentions ou des sentiments qui dépassent ses capacités réelles, ce qui interroge la frontière entre relation éducative et relation affective avec un artefact technologique.

De plus, les mécanismes sous-jacents à cette interaction émotionnelle restent encore partiellement compris, nécessitant des recherches complémentaires pour en cerner les déterminants, la stabilité dans le temps et l'impact sur le développement émotionnel et social des enfants.

Enfin, il conviendrait d'approfondir la réflexion sur les effets à long terme de l'introduction de tels agents artificiels dans des pratiques éducatives, afin de ne pas surévaluer leur potentiel au détriment d'une approche critique et humaniste de la relation éducative. La technologie ne doit pas se substituer à l'humain, mais plutôt s'inscrire comme un outil médiateur au service de la pédagogie, de l'apprentissage et du développement global de l'enfant.

## 4.2 Expérience 2 : Apprendre avec Buddy à l'INSPE d'Aix-Marseille

L'expérience menée à l'INSPE d'Aix-Marseille ainsi que dans d'autres instances d'Aix-Marseille Université s'inscrit dans un cadre pédagogique visant à intégrer des technologies innovantes, telles que la robotique éducative, au sein des cursus destinés aux futurs professeurs des écoles, de collège et de lycée. Ce projet a également impliqué des ingénieurs pédagogiques et des élèves de lycées. L'objectif est d'explorer comment des robots comme **Buddy**, en combinaison avec des outils d'IA générative, peuvent enrichir les pratiques pédagogiques et faciliter l'apprentissage des compétences technologiques chez les étudiants et les élèves de manière interactive et engageante.

TABLE 4 – Répartition des participant·e·s à l'étude sur l'usage pédagogique de **Buddy**

| **Niveau** | **Nombre de participant·e·s** |
|---|---|
| Lycéen.n.es | 79 |
| Étudiant.e.s de l'INSPE d'Aix-Marseille | 59 |
| Autres étudiant.e.s | 23 |
| **Total** | **161 participant.e.s** |

Le tableau 4 ci-dessus présente la répartition des participant·e·s ayant pris part à l'étude observationnelle menée dans le cadre de l'analyse de l'usage pédagogique de **Buddy**, un agent conversationnel utilisé en contexte éducatif. Cette étude a mobilisé un échantillon diversifié composé de lycéen·ne·s, d'étudiant·e·s en formation initiale à l'INSPE d'Aix-Marseille, ainsi que d'autres étudiant·e·s issu·e·s de formations universitaires variées. La diversité des profils vise à permettre une analyse comparative des usages et perceptions de l'agent en fonction du niveau de formation des participant·e·s.

### 4.2.1 Quels usages pédagogiques ?

**a. Prise de conscience technologique et conception de ressources pédagogiques.** Un des premiers usages de **Buddy** a été d'aider les étudiants à prendre conscience du potentiel de la robotique éducative pour la conception de ressources pédagogiques. En manipulant **Buddy**, les étudiants ont pu imaginer des applications concrètes de la robotique dans l'enseignement des matières scientifiques, notamment en ce qui concerne le codage, la logique algorithmique, et la modélisation des systèmes.

**b. Initiation des jeunes enfants au codage à travers un robot attractif.** L'une des principales applications de **Buddy** dans cette expérience était d'initier les enfants au codage et au raisonnement (par exemple, tracer une trajectoire) dès le plus jeune âge. Le robot, avec son apparence inspirée des dessins animés et sa voix robotisée, est particulièrement adapté pour des enfants qui peuvent être intimidés par des robots à l'apparence trop réaliste. **Buddy** est conçu pour être à la fois rassurant et engageant, offrant une interface ludique permettant aux jeunes élèves de s'initier à des concepts complexes de manière progressive et divertissante.

**c. Buddy comme assistant intelligent pour l'engagement des étudiants. Buddy**, en intégrant un modèle de langage large (LLM), a également joué un rôle crucial en tant qu'assistant interactif pour les étudiants. En répondant à leurs questions de manière contextuelle et personnalisée, le robot a favorisé l'engagement, la motivation et la participation active des étudiants. Cette interaction, dynamique et basée sur la conversation, a permis de maintenir l'intérêt des étudiants tout au long des sessions d'apprentissage, tout en facilitant leur compréhension des concepts pédagogiques.

**d. Démythification de l'IA et des limites de la technologie.** Enfin, un aspect fondamental de l'utilisation de **Buddy** a été de mettre en lumière les limites techniques du robot et de l'IA, en expliquant les bases techniques d'un modèle de langage large (LLM). En effet, malgré les avancées technologiques, **Buddy** présente des lacunes qui rendent la robotique et l'IA accessibles à une analyse critique. Ces imperfections ont servi d'exemple concret pour illustrer les contraintes actuelles de l'IA, en particulier en matière de compréhension contextuelle, de capacités de raisonnement complexes et d'interactions authentiques.

### 4.2.2 Résultats, discussion et limites

Cette section propose une analyse des données préliminaires recueillies dans le cadre de l'étude d'observation sur l'utilisation du robot **Buddy** en contexte pédagogique. Les résultats sont présentés selon plusieurs critères afin de mettre en évidence les tendances observées et d'enrichir la compréhension des modalités d'usage de cet agent robotique par les différents publics concernés.

**a. Facilité d'utilisation (71%).** Le fait que 71% des utilisateurs trouvent le robot facile à utiliser indique une adoption plutôt positive, mais il y a encore une marge d'amélioration. Ce chiffre suggère qu'une partie des utilisateurs pourrait rencontrer des difficultés, soit dans la prise en main, soit dans l'intégration du robot dans un environnement d'apprentissage. Il serait pertinent de réaliser des analyses qualitatives pour mieux comprendre les obstacles rencontrés par les utilisateurs.

**b. Appréciation de la posture humanoïde et du visage expressif (92%).** Une très forte majorité des utilisateurs (92%) semble apprécier l'aspect humanoïde du robot ainsi que son visage expressif. Cette caractéristique pourrait jouer un rôle clé dans l'engagement émotionnel des apprenants, rendant le robot plus accessible et stimulant pour les interactions.

**c. Fluidité de l'interaction vocale avec Buddy (57%).** Le score de 57% pour la fluidité de l'interaction vocale pourrait refléter certaines limitations dans la reconnaissance ou la génération de la parole par le robot. Une réévaluation de la qualité du micro et/ou du traitement du langage naturel (TLN) et de l'intelligibilité des réponses pourrait être nécessaire pour améliorer cette interaction.

**d. Utile pour l'apprentissage des langues (88%).** Le robot semble avoir une grande valeur pédagogique dans l'enseignement des langues, avec 88% des utilisateurs considérant l'outil utile dans ce domaine. Cela peut être dû à la capacité du robot à pratiquer des conversations, corriger des erreurs de prononciation et offrir un retour immédiat.

**e. Utile pour l'apprentissage du codage (79%).** Bien que légèrement inférieur, le score de 79% indique également une bonne utilité pour l'apprentissage du codage. **Buddy** pourrait être utilisé pour des activités telles que l'introduction aux concepts de programmation ou la simulation de certains processus informatiques dans des environnements interactifs.

**f. Engagement des apprenants (93%).** Un taux d'engagement aussi élevé (93%) montre l'impact positif de l'outil en termes de motivation et d'interaction. L'usage de robots dans les classes semble jouer un rôle crucial dans la stimulation de l'attention et de la participation des étudiants.

**g. Recommandation de la robotique dans les pratiques pédagogiques (86%).** Le fait que 86% des répondants recommandent la robotique en pédagogie souligne l'efficacité perçue de cette approche innovante. L'intégration de la robotique en classe semble être bien reçue par les enseignants et les apprenants, probablement en raison de l'aspect ludique et interactif.

**h. Activités avec Buddy.** Les activités proposées, telles que l'outil conversationnel, la vérification de la prononciation, le feedback instantané, la traduction, et la communication avec des élèves allophones, montrent la polyvalence de **Buddy** dans divers contextes pédagogiques. De plus, l'inclusion des élèves éloignés (par exemple, hospitalisés) montre le potentiel de **Buddy** pour favoriser une inclusion éducative élargie avec un coût relativement acceptable mais une volonté institutionnelle est nécéssaire.

**i. Avis des élèves sur Buddy.** Les élèves ont une perception de **Buddy** comme :

- Mignon et attractif : Ce côté ludique semble bien fonctionner pour attirer l'attention des élèves.
- Attachement émotionnel : L'aspect humanoïde et l'interaction affective contribuent probablement à un lien plus fort avec l'élève.
- Ludique et amusant : L'aspect de jeu est souvent un facteur important dans l'engagement des élèves, notamment pour ceux qui apprennent de manière plus pratique.
- Découverte du monde de la robotique et de l'IA : Les élèves peuvent être inspirés par l'opportunité d'interagir avec une technologie avancée, ce qui pourrait éveiller leur curiosité pour la robotique et l'intelligence artificielle.

Dans l'ensemble, les retours semblent indiquer que **Buddy** est un outil pédagogique prometteur, notamment pour l'apprentissage des langues et l'engagement des élèves. Cependant, il y a des domaines à améliorer, notamment la fluidité de l'interaction vocale et la facilité d'utilisation générale. L'approche centrée sur l'interaction ludique et émotionnelle semble jouer un rôle clé dans son efficacité, en particulier pour l'attachement des élèves et leur motivation.

À l'issue de cette étude, plusieurs recommandations peuvent être formulées en vue d'optimiser l'intégration pédagogique de **Buddy**. Tout d'abord, l'amélioration de l'interaction vocale constitue une priorité : le renforcement des performances de la reconnaissance vocale et de la qualité des réponses du robot permettrait de fluidifier les échanges et de limiter les obstacles à la communication. Par ailleurs, il apparaît essentiel de proposer des sessions de formation destinées aux enseignant·e·s, afin de les accompagner dans l'appropriation de l'outil et son intégration dans leurs pratiques pédagogiques. Enfin, une plus grande personnalisation des interactions et des activités proposées par **Buddy**, en fonction des besoins spécifiques des élèves, tels que le niveau de compétence ou les centres d'intérêt, contribuerait à renforcer la pertinence pédagogique de l'agent robotique.

# 5 Conclusion

L'exploration de l'intégration de robots comme **Buddy** dans les pratiques pédagogiques interroge autant qu'elle fascine. Si ces dispositifs, dépourvus d'émotions biologiques, ne peuvent prétendre à une authenticité affective, leur capacité à susciter des réponses émotionnelles et réflexives chez l'humain ouvre des perspectives éducatives inédites. En incarnant un médiateur technologique doté d'expressivité, **Buddy** transcende le statut d'outil fonctionnel pour devenir un catalyseur d'interactions, stimulant l'introspection, la créativité et le lien social. Son potentiel réside moins dans la simulation d'émotions que dans sa faculté à révéler, par contraste, la complexité et la richesse des dynamiques humaines.

Les limites techniques actuelles, telles que la répétitivité des expressions, rappellent que ces robots sont des partenaires en évolution, invitant à renforcer les collaborations entre ingénierie, sciences cognitives, arts et pédagogie. Loin de marginaliser le rôle de l'enseignant, cette technologie le recentre comme pilier éthique et créateur de sens, garantissant que l'usage de la machine reste ancré dans une finalité émancipatrice. En associant médiation artistique, réflexivité et cadre déontologique, **Buddy** illustre une vision de l'éducation où la technologie n'aliène pas, mais enrichit le dialogue entre l'élève, le formateur et l'IA générative enrichie par de multiples sources de données et confirmations issues de notre monde avec toutes ses dimensions (culturelles, politiques, etc.).

Cet article invite ainsi à repenser la pédagogie non comme un champ à automatiser, mais comme un espace de résonance où la machine, en miroir de nos émotions, interroge notre humanité et notre intelligence. L'enjeu n'est pas de rivaliser avec l'humain, mais de cultiver, à travers ces nouvelles interfaces, une relation éducative plus empathique, inclusive et consciente de ses propres enjeux. L'avenir de la robotique éducative se dessine alors dans cet équilibre exigeant : faire de la technologie un allié pour mieux célébrer ce qui nous rend irremplaçables, la rendre au service de l'enseignement, de l'apprentissage, de la pédagogie et pas l'inverse.

# Références


ANZALONE S. M., BOUCENNA S., IVALDI S. & CHETOUANI M. (2015). Evaluating the engagement with social robots. *International Journal of Social Robotics*, **7**, 465–478. DOI : 10.1007/s12369-015-0298-7.

BADACHE I. & BELLET P. (2024). Intelligence artificielle : usage pédagogique et esprit critique. In *16ème édition du colloque Interactions Multimodales Par ÉCran, IMPEC 2024*. https://hal.science/hal-04659335v1.

BARAKOVA E., VÄÄNÄNEN K., KAIPAINEN K. & MARKOPOULOS P. (2023). Benefits, challenges and research recommendations for social robots in education and learning : A meta-review. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, p. 2555–2561. DOI : 10.1109/RO-MAN57019.2023.10309345.

BELPAEME T., KENNEDY J., RAMACHANDRAN A., SCASSELLATI B. & TANAKA F. (2018). Social robots for education : A review. *Science Robotics*, **3**(21), eaat5954. DOI : 10.1126/scirobotics.aat5954.

BELPAEME T. & TANAKA F. (2021). Social robots as educators. In *OECD Digital Education Outlook 2021 : Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots*, p. 143. OECD Publishing Paris. DOI : 10.1787/589b283f-en.

BREAZEAL C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, **42**(3), 167–175. Socially Interactive Robots, DOI : 10.1016/S0921-8890(02)00373-1.

BREAZEAL C. (2004a). *Designing sociable robots*, In *Designing Sociable Robots*, p. 27–50. MIT press. http://ieeexplore.ieee.org/document/6280040.

BREAZEAL C. (2004b). Social interactions in HRI : the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **34**(2), 181–186. DOI : 10.1109/TSMCC.2004.826268.

BREAZEAL C., DAUTENHAHN K. & KANDA T. (2016). *Social Robotics*, In B. SICILIANO & O. KHATIB, Éds., *Springer Handbook of Robotics*, p. 1935–1972. Springer International Publishing : Cham. DOI : 10.1007/978-3-319-32552-1_72.

FONG T., NOURBAKHSH I. & DAUTENHAHN K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, **42**(3), 143–166. Socially Interactive Robots, DOI : https ://doi.org/10.1016/S0921-8890(02)00372-X.

HEGEL F., MUHL C., WREDE B., HIELSCHER-FASTABEND M. & SAGERER G. (2009). Understanding social robots. In *2009 Second International Conferences on Advances in Computer-Human Interactions*, p. 169–174. DOI : 10.1109/ACHI.2009.51.

JOHAL W. (2020). Research trends in social robots for learning. *Current Robotics Reports*, **1**(3), 75–83. DOI : 10.1007/s43154-020-00008-3.

KIRBY R., FORLIZZI J. & SIMMONS R. (2010). Affective social robots. *Robotics and Autonomous Systems*, **58**(3), 322–332. Towards Autonomous Robotic Systems 2009 : Intelligent, Autonomous Robotics in the UK, DOI : 10.1016/j.robot.2009.09.015.

LEITE I., MARTINHO C. & PAIVA A. (2013). Social robots for long-term interaction : a survey. *International Journal of Social Robotics*, **5**, 291–308. DOI : 10.1007/s12369-013-0178-y.

MAHDI H., AKGUN S. A., SALEH S. & DAUTENHAHN K. (2022). A survey on the design and evolution of social robots — past, present and future. *Robotics and Autonomous Systems*, **156**, 104193. DOI : 10.1016/j.robot.2022.104193.

MILLIEZ G. (2018). Buddy : A companion robot for the whole family. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, p.40, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3173386.3177839.

MORI M., MACDORMAN K. F. & KAGEKI N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, **19**(2), 98–100. DOI : 10.1109/MRA.2012.2192811.

PACHIDIS T., VROCHIDOU E., KABURLASOS V. G., KOSTOVA S., BONKOVIĆ M. & PAPIĆ V. (2019). Social robotics in education : State-of-the-art and directions. In N. A. ASPRAGATHOS, P. N. KOUSTOUMPARDIS & V. C. MOULIANITIS, Éds., *Advances in Service and Industrial Robotics*, p. 689–700, Cham : Springer International Publishing. DOI : 10.1007/978-3-030-00232-9_72.

PAPERT S. (1980). *Mindstorms : children, computers, and powerful ideas*. USA : Basic Books, Inc. https://dl.acm.org/doi/book/10.5555/1095592.

SMAKMAN M., VOGT P. & KONIJN E. A. (2021). Moral considerations on social robots in education : A multi-stakeholder perspective. *Computers & Education*, **174**, 104317. DOI : 10.1016/j.compedu.2021.104317.

TANG X. & CHU J. (2022). Inclusive design : Task specified robots for elderly. *Advances in Education, Humanities and Social Science Research*, **1**(1), 82–82. DOI : 10.56028/aehssr.1.1.82.

UNESCO (2021). *Reimagining our futures together : A new social contract for education*. Paris : United Nations Educational, Scientific and Cultural Organization. Report of the International Commission on the Futures of Education. https://unesdoc.unesco.org/ark:/48223/pf0000379381.

VAN DEN BERGHE R., VERHAGEN J., OUDGENOEG-PAZ O., VAN DER VEN S. & LESEMAN P. (2019). Social robots for language learning : A review. *Review of Educational Research*, **89**(2), 259–295. DOI : 10.3102/0034654318821286.

WOO H., LETENDRE G. K., PHAM-SHOUSE T. & XIONG Y. (2021). The use of social robots in classrooms : A review of field-based studies. *Educational Research Review*, **33**, 100388. DOI : 10.1016/j.edurev.2021.100388.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# SEPT : Détecter les difficultés des étudiants à travers le clustering de leurs trajectoires émotionnelles et physique lors d'évaluations en ligne sur Moodle

Edouard Nadaud[1, 2] Antoun Yaacoub[1] Bénédicte Legrand[2] Lionel Prevost[1, 2]
(1) ESIEAlab, équipe Learning Data Robotic (LDR), ESIEA, 74 bis Av. Maurice Thorez, 94200 Ivry-sur-Seine, France
(2) Centre de Recherche en Informatique (CRI), Paris 1 panthéon Sorbonne, 31, rue Baudricourt 75013 Paris, France
edouard.nadaud@esiea.fr, antoun.yaacoub@esiea.fr,
benedicte.le-grand@univ-paris1.fr, lionel.prevost@esiea.fr

## RÉSUMÉ

Imaginez une salle de classe où les difficultés et réussites des étudiants s'expriment non par des mots, mais par l'expression de leurs visages et mouvements, captés en temps réel pendant un quiz. Les méthodes d'enseignement dans le supérieur se font de plus en plus hybride et à distance. Les interactions directes sont réduites, rendant difficile la détection des moments de décrochage. Pour y remédier, nous introduisons le concept de Trajectoires Émotionnelles et Physiques Étudiantes (SEPT). Grâce aux webcams de 89 étudiants de première année de Master, nous avons enregistré et analysé chaque seconde leurs expressions faciales (valence, arousal selon le modèle de Russell) et états physiques (orientation de la tête, distance à l'écran). Les séries temporelles ainsi obtenues révèlent des motifs distincts selon que les difficultés soient individuelles ou liées aux questions. SEPT offres des perspectives pour des systèmes intelligents de suivi affectif en contexte éducatif numérique.

## ABSTRACT

**SEPT : Uncovering Student Difficulties through Emotional and Physical Trajectories during Online Assessments**

Imagine a classroom where students' difficulties and successes are expressed not by words, but by their facial expressions and movements, captured in real time during a quiz. Teaching methods in higher education are increasingly hybrid and remote. Direct interactions are reduced, making it difficult to detect moments of disengagement. To address this, we introduce the concept of Student Emotional and Physical Trajectories (SEPT). Using the webcams of 89 first year Master's students, we recorded and analyzed every second their facial expressions (valence, arousal according to the Russell model) and physical states (head orientation, distance from the screen). The resulting time series reveal distinct patterns depending on whether the difficulties are individual, or question related. SEPT offers perspectives for intelligent systems for affective monitoring in a digital educational context.

MOTS-CLÉS : Informatique affective, trajectoires émotionnelles, applications de l'apprentissage automatique, IA éthique, analytique de l'apprentissage..

KEYWORDS: Affective Computing, Emotion Trajectories, Machine-Learning Applications, Ethical AI, Learning analytics..

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 111–125.

# 1 Introduction

Les interactions traditionnelles en présentiel entre enseignants et étudiants deviennent moins fréquentes à l'ère de l'enseignement hybride et à distance (Shi *et al.*, 2024). Cette évolution offre certes une flexibilité accrue et un accès plus large à l'apprentissage, mais soulève une préoccupation majeure : comment prévenir l'abandon scolaire lorsque les enseignants ne peuvent pas directement observer les signes de difficultés chez les étudiants ? Sans indices visuels et interpersonnels immédiats, il devient difficile pour les enseignants de détecter lorsque les étudiants sont confus, frustrés ou désengagés (Frenkel *et al.*, 2024). Des recherches antérieures ont établi que les difficultés académiques sont souvent liées aux états émotionnels des étudiants (Frenkel *et al.*, 2024). Notre approche s'appuie sur des travaux précédents qui soulignent les limites des mesures émotionnelles statiques et agrégées pour prédire les résultats académiques (Nadaud *et al.*, 2024). Nous proposons une nouvelle approche nommée « Trajectoires Émotionnelles et Physiques des Étudiants (SEPT) », afin de surveiller continuellement les signaux émotionnels et comportementaux des étudiants durant les activités d'apprentissage. L'objectif est de réintroduire une forme de rétroaction « visuelle » pour les enseignants dans les environnements d'apprentissage à distance, permettant ainsi d'identifier rapidement les difficultés des étudiants et d'offrir un soutien en temps opportun. À partir de ces motivations, nous formulons les hypothèses clés suivantes :

**H1. La valence et l'arousal seuls sont insuffisants :** Bien que la valence et l'arousal offrent une mesure basique de l'affect, elles ne capturent pas complètement l'état émotionnel de l'étudiant. Nous supposons que l'intégration d'états physiques additionnels (comme les mouvements physiques, faciaux et la direction du regard) améliore la précision et la richesse de l'évaluation émotionnelle.

**H2. Les émotions varient à l'échelle des questions individuelles :** L'état émotionnel n'est pas statique tout au long d'une évaluation, mais fluctue dynamiquement au sein de chaque question du quiz. Capturer ces variations fines est essentiel pour comprendre en temps réel la frustration et l'état mental des étudiants.

**H3. Les trajectoires émotionnelles dépendent du contexte, et ne sont pas constantes pour un étudiant donné :** Nous faisons l'hypothèse qu'un même étudiant peut présenter des réponses émotionnelles très différentes selon la charge cognitive de chaque question. Cela remet en question l'idée d'un état émotionnel stable propre à l'étudiant, suggérant ainsi la nécessité de considérer le contexte.

**H4. Existence de modèles communs pour les questions difficiles :** Certains modèles SEPT pourraient réapparaître chez différents étudiants confrontés à une même question difficile. Autrement dit, des défis cognitifs similaires pourraient provoquer des trajectoires SEPT comparables chez plusieurs étudiants, révélant ainsi des points de difficulté communs.

Pour tester ces hypothèses, nous avons développé une approche capturant continuellement les expressions faciales et les mouvements physiques et faciaux des étudiants via leurs webcams durant les quiz. Notre système suit les états émotionnels via les métriques de valence et d'arousal, et intègre des indicateurs physiques afin d'offrir un portrait complet et temporellement détaillé du comportement des étudiants. En examinant les émotions et les comportements au niveau granulaire des questions individuelles, nous cherchons à identifier les moments critiques de difficulté, fournissant ainsi des informations susceptibles de permettre des interventions pédagogiques opportunes. Finalement, notre approche réintroduit un élément crucial de l'expérience en classe : la rétroaction émotionnelle en temps réel dans les contextes d'apprentissage hybrides et à distance en exploitant les progrès de

l'informatique affective. Ce travail rapproche l'informatique affective de la pédagogie, offrant aux enseignants des outils pour comprendre et soutenir les apprenants d'une manière jusqu'alors impossible dans les environnements distants.

La suite de l'article est organisée comme suit : la Section 2 présente les travaux connexes, la Section 3 décrit notre méthodologie, la Section 4 expose les résultats obtenus, et enfin la conclusion traite les limites de notre approche et propose des pistes pour des recherches futures.

## 2 Travaux connexes

Au cours des dernières années, les interactions entre les émotions des étudiants, leurs indicateurs physiques et leurs performances académiques ont suscité un intérêt croissant, notamment en raison du passage des salles de classe traditionnelles vers des environnements d'apprentissage en ligne et hybrides (Spitzer & Moeller, 2023).

**Évolution des environnements d'apprentissage :** L'expansion rapide des plateformes d'apprentissage numérique, particulièrement accélérée par la pandémie de COVID-19, a considérablement remodelé l'enseignement supérieur (Chaubey & Bhattacharya, 2015). Garrison *et al.* (Garrison *et al.*, 1999) soulignent qu'un apprentissage efficace nécessite la présence cognitive, pédagogique et sociale ; cette dernière étant souvent réduite dans les contextes à distance. Cette perte d'interaction directe peut créer une « distance émotionnelle » entre enseignants et étudiants, rendant plus difficile pour les enseignants la détection des difficultés rencontrées par les étudiants. Blikstein et Worsley (Blikstein & Worsley, 2016) précisent que ce manque de rétroaction est particulièrement problématique lors des évaluations, où les possibilités d'intervention pédagogique en temps réel sont limitées. Ces défis ont encouragé les chercheurs à explorer des méthodes technologiques permettant de combler l'écart entre la perception des enseignants et le soutien aux étudiants.

**Émotions et résultats d'apprentissage :** De nombreuses preuves démontrent que les émotions jouent un rôle central dans les résultats d'apprentissage. Par exemple, D'Mello et Graesser (Graesser & D'Mello, 2012) ont montré que les états affectifs transitoires tels que l'engagement, la confusion et la frustration peuvent influencer directement le traitement cognitif et la rétention des connaissances. Les états émotionnels positifs sont associés à une plus grande motivation et au plaisir d'apprendre (Harley *et al.*, 2015), alors que les états négatifs comme l'ennui ou l'anxiété peuvent entraver les progrès d'apprentissage (Arroyo *et al.*, 2014). Ces observations ont encouragé l'intégration de l'informatique affective dans les environnements éducatifs, dans le but de suivre les expériences émotionnelles des étudiants et d'y répondre de manière à améliorer l'apprentissage.

**Détection des émotions et trajectoires en contexte éducatif :** Les premières méthodes de détection des émotions des étudiants reposaient sur des auto-évaluations ou des évaluations par des observateurs (Baker *et al.*, 2010). Ces méthodes ont été critiquées en raison de leur subjectivité et du risque de perturber le processus d'apprentissage. Les progrès dans la détection sensorielle et la vision par ordinateur ont permis d'adopter des approches plus objectives. Des capteurs physiologiques (par exemple, des moniteurs de fréquence cardiaque ou des dispositifs de conductance cutanée) ont été utilisés pour mesurer l'arousal et les niveaux de stress (Jiang *et al.*, 2018), et des techniques basées sur des caméras peuvent suivre le regard et les expressions faciales pour inférer l'attention et les émotions (Arroyo *et al.*, 2014; Bosch *et al.*, 2016). Cependant, beaucoup de ces approches nécessitent un équipement spécialisé ou des conditions contrôlées, limitant ainsi leur extensibilité dans des

salles de classe réelles (Paquette *et al.*, 2014). Pour modéliser les émotions détectées, la plupart des études précédentes utilisaient soit les catégories émotionnelles discrètes d'Ekman (Perry, 2014), soit traitaient l'affect selon des dimensions continues moyennées sur une activité (par exemple, valence et arousal moyens) (Jiang *et al.*, 2018). (La valence réfère au caractère positif ou négatif de l'émotion, et l'arousal correspond au niveau d'activation ou d'alerte.) Des recherches récentes par Harley *et al.* (Harley *et al.*, 2017) et Silva *et al.* (Silva *et al.*, 2014) affirment que ces représentations statiques ne parviennent pas à capturer les fluctuations rapides et contextuelles des émotions pendant les tâches d'apprentissage. Ceci a conduit à un intérêt croissant pour l'analyse temporelle des données affectives, examinant comment les émotions évoluent au fil du temps plutôt qu'en considérant seulement des moyennes globales. Dans une étude antérieure (Nadaud *et al.*, 2024), une émotion par question, représentant le centroïde de toutes les valeurs émotionnelles enregistrées pendant la réponse à cette question, a été calculée. Ces centroïdes ont été connectés pour former des trajectoires tout au long du quiz. Les auteurs avaient émis l'hypothèse que des notes faibles correspondraient à des centroïdes émotionnels négatifs, mais cette méthode n'a pas montré de corrélation significative entre les états émotionnels moyennés et les performances.

**Intégration d'indicateurs physiques :** Les expressions faciales seules offrent une vue partielle de l'état des étudiants. Les états physiques tels que la posture et les mouvements reflètent également l'effort cognitif (Arroyo *et al.*, 2014), et leur intégration avec les données faciales peut améliorer la prédiction de la frustration et des performances (Whitehill *et al.*, 2014; Bosch & D'Mello, 2017). Les systèmes multimodaux surpassent systématiquement les approches unimodales (D'mello & Kory, 2015), bien que la fusion temporelle de ces signaux reste difficile en raison de leur asynchronisme.

**Techniques avancées d'analyse temporelle :** Des études récentes ont appliqué des méthodes analytiques basées sur les séquences aux données étudiantes, révélant des modèles comportementaux nuancés. Par exemple, Moreno-Marcos *et al.* (Moreno-Marcos *et al.*, 2020) ont utilisé le clustering pour identifier des trajectoires distinctes d'engagement dans des cours en ligne, constatant que certaines trajectoires temporelles d'interaction sont associées à de meilleurs taux d'achèvement. Rodrigo *et al.* (Rodrigo *et al.*, 2012) ont utilisé l'analyse de séquences sur les états affectifs des étudiants (par exemple, confusion $\rightarrow$ insight $\rightarrow$ confusion), et ont associé des séquences émotionnelles spécifiques à des gains d'apprentissage ou au désengagement. Piot *et al.* (Piot *et al.*, 2019) explorent « l'effet eureka » (confusion $\rightarrow$ surprise $\rightarrow$ joie). Zhou *et al.* (Zhao & Itti, 2016) ont introduit l'utilisation du Dynamic Time Warping (DTW) pour aligner et comparer les séquences temporelles des comportements des apprenants, en tenant compte des différences individuelles de rythme. Malgré ces progrès, des lacunes importantes subsistent. Paquette *et al.* (Paquette *et al.*, 2014) ont noté que peu d'efforts intégraient les indicateurs physiques (comme la posture corporelle ou le regard) avec les données émotionnelles en une trajectoire multidimensionnelle unique.

En s'appuyant sur une première exploration des trajectoires émotionnelles (Nadaud *et al.*, 2024), cet article présente les SEPT comme un avancement complet. Les SEPT étendent le modèle bidimensionnel valence-arousal à un cadre à sept dimensions, incorporant des indicateurs physiques (orientation de la tête, distance du visage à l'écran) en complément des caractéristiques émotionnelles. Les SEPT capturent des fluctuations continues, seconde par seconde, en exploitant le DTW et le clustering afin d'identifier des motifs parmi les trajectoires à un niveau de granularité fin.

# 3 Méthodologie

Pour tester nos hypothèses, nous avons mené une expérimentation contrôlée en plusieurs phases. Nous décrivons ci-dessous précisément le cadre expérimental et chaque étape de notre méthodologie.

## 3.1 Cadre expérimental

Nous avons réalisé notre étude auprès de 89 étudiants de première année de Master dans une école d'ingénieurs française. Tous les participants avaient une formation en technologies de l'information et étaient à l'aise avec l'utilisation d'ordinateurs portables et de webcams. L'expérience s'est déroulée durant un cours en présentiel où chaque étudiant utilisait son propre ordinateur portable muni d'une webcam intégrée afin de passer un quiz en ligne. Afin de maintenir un environnement naturel, aucune consigne spécifique concernant leur posture ou la position de la caméra n'a été donnée ; les étudiants interagissaient normalement avec le quiz. Le quiz, proposé au début d'un cours sur l'apprentissage automatique, comportait 6 à 7 questions de différents formats : tâches de codage, questions à choix multiples et exercices de type glisser-déposer. Chaque question ne permettait qu'une seule tentative, sans retour en arrière, assurant une progression linéaire. Les étudiants disposaient de 12 minutes au maximum pour terminer toutes les questions, après quoi le quiz se fermait automatiquement et était noté. Durant le quiz, les étudiants étaient tenus de rester sur la page sans possibilité de navigation externe. Nous avons utilisé l'extension Moodle Proctoring (adaptée de sa fonctionnalité initiale de vérification d'identité) afin de capturer des images de chaque étudiant à intervalles réguliers tout au long du quiz. Ce dispositif a permis une observation continue des expressions faciales et comportements des étudiants tout en préservant une expérience naturelle de passation de quiz.

## 3.2 Collecte des données

Notre dispositif expérimental réduit les besoins matériels à un simple ordinateur portable équipé d'une webcam, évitant ainsi intentionnellement l'utilisation de dispositifs coûteux tels que des bracelets capteurs ECG/EDM (Spitzer & Moeller, 2023; Nandi *et al.*, 2021). Les données collectées consistent en des images capturées toutes les secondes. L'interface utilisée est strictement identique à celle d'un quiz standard sans surveillance vidéo, préservant ainsi une expérience utilisateur fluide et minimisant toute influence sur les performances au quiz (Lee, 2020). Simultanément, nous avons collecté des données complètes provenant du système de gestion de l'apprentissage (LMS), incluant les notes obtenues, le temps passé par question, et les niveaux de difficulté des questions.

## 3.3 Conformité RGPD et considérations éthiques de l'étude

Cette étude a été menée en conformité stricte avec les directives éthiques et le Règlement Général sur la Protection des Données (RGPD) (Union, 2018) afin de protéger la vie privée des participants. Plusieurs mesures ont été mises en œuvre pour garantir l'éthique de la recherche :

— **Consentement éclairé :** La participation était entièrement volontaire. Les étudiants ont été pleinement informés des objectifs du projet, des données collectées (images périodiques de la webcam et données de performance) et de leur utilisation exclusive à des fins de recherche. Un consentement écrit explicite a été obtenu auprès de chaque participant avant la collecte

des données. Les étudiants étaient libres de refuser ou de se retirer de l'étude à tout moment, sans aucune conséquence académique.

— **Préparation et transparence :** Avant le quiz noté, un quiz d'entraînement a été organisé avec le même dispositif de surveillance pour que les participants puissent s'y familiariser. Les étudiants refusant de participer passaient le quiz dans les mêmes conditions à l'exception de la capture vidéo, préservant ainsi l'équité des conditions d'examen.
— **Approbation éthique :** Le protocole de recherche a été revu et approuvé par le comité d'éthique de l'école, garantissant la conformité avec les standards éthiques établis pour la recherche impliquant des participants humains.
— **Conformité juridique et sécurité des données :** Nous avons consulté un avocat spécialisé en protection des données pour concevoir nos formulaires de consentement et notre plan de gestion des données, en conformité totale avec le RGPD. Toutes les données collectées (images et données LMS) ont été anonymisées et stockées sur un serveur sécurisé de l'école, accessible uniquement par l'équipe de recherche autorisée.
— **Préservation de la confidentialité :** Les images webcam ont été utilisées exclusivement pour extraire automatiquement les expressions faciales et les mouvements de la tête, sans reconnaissance d'identité.

## 3.4 Prétraitement des données

Avant l'analyse émotionnelle, nous avons appliqué une série d'étapes de prétraitement afin d'améliorer la qualité des données. En conditions réelles de classe, certaines images capturées peuvent être impropres à la reconnaissance émotionnelle (visage partiellement visible, mauvaise luminosité, etc.). Un pipeline de filtrage a été implémenté pour ne conserver que des images exploitables. Après avoir évalué plusieurs algorithmes de détection faciale (Haar Cascade, dlib, MTCNN, YOLOv5), nous avons retenu le modèle YOLOv5 (Ieamsaard *et al.*, 2021) pour sa précision et sa robustesse. Les images sans visage détecté ont subi une amélioration (correction gamma, égalisation d'histogramme), suivie d'une seconde tentative de détection. Les images restantes non valides ont été exclues. Ce processus a conduit à la sélection finale de 82 étudiants (sur les 89 initiaux), assurant ainsi une meilleure qualité des données.

## 3.5 Méthodes d'analyse des données

Après nettoyage, nous avons analysé les SEPT via des techniques de vision par ordinateur et apprentissage automatique. L'analyse comportait deux étapes principales : (1) prédiction continue des états émotionnels (valence, arousal) à partir des images, et (2) intégration avec des indicateurs physiques pour former les SEPT.

**Prédiction des émotions.** Pour inférer avec précision les états affectifs à partir d'environ 99 000 images capturées et les cartographier selon le modèle d'affect de Russell, nous avons abandonné les méthodes d'annotation traditionnelles en raison de leur nature gourmande en ressources et de leur subjectivité, qui entraînent souvent une faible concordance entre annotateurs et des délais prolongés (Graesser & D'Mello, 2012). Le volume et la complexité de notre ensemble de données d'images ont rendu les approches conventionnelles d'apprentissage automatique inadéquates, car elles peinent à traiter des données de grande dimension sans traitement manuelle significative des caractéristiques. Nous avons adopté une approche d'apprentissage profond pour surmonter les limites

des modèles de prédiction des émotions 2D existants. En exploitant les ensembles de données AffectNet (Mollahosseini *et al.*, 2019) et Affwild (Liu & Kollias, 2019), largement reconnus et couramment utilisés dans la recherche en informatique affective, nous avons entraîné deux modèles à prédire la valence et l'arousal.

| *Modèle* | **Valence** | | **Arousal** | |
|---|---|---|---|---|
| | *CCC* | *RMSE* | *CCC* | *RMSE* |
| CAGE 2024 (Wagner *et al.*, 2024) | **0.716** | **0.331** | 0.642 | **0.305** |
| VGG-F 2021 (Bulat *et al.*, 2022) | 0.710 | 0.356 | 0.629 | 0.326 |
| Nos modèles | 0.714 | 0.339 | **0.644** | 0.323 |

TABLE 1 – Comparaison des performances des modèles sur les dimensions de valence et d'arousal.

Le tableau 1 compare nos modèles aux données de pointe en matière de prédiction de l'arousal de valence. Chaque modèle comprend cinq couches convolutives et de sous-échantillonnage séquentiel (Conv2D, Conv2D, MaxPooling), suivies de deux couches denses, totalisant environ 2 millions de paramètres. Cette architecture a permis des prédictions émotionnelles précises, avec une perte de validation inférieure à 0,1, permettant une représentation détaillée des états émotionnels des étudiants, conforme au modèle de Russell (voir Fig. 1 - droite). La représentation circulaire 2D obtenue capture efficacement les expressions faciales actives, les expressions les plus prononcées étant positionnées plus à droite ou à gauche du cercle.

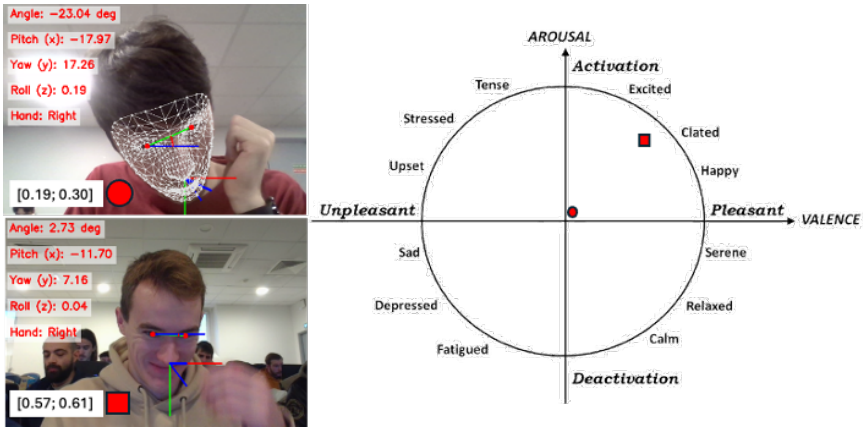



FIGURE 1 – Exemple de capture de données 7D pour la construction de la trajectoire SEPT

**SEPT.** Outre la valence et l'arousal, nous avons extrait l'orientation tridimensionnelle de la tête (tangage, lacet et roulis) et une estimation de la distance de l'étudiant à l'écran pour chaque image (basée sur la taille du cadre de délimitation du visage comme indicateur). Ainsi, chaque image est représentée par un vecteur de caractéristiques à 7 dimensions : [valence, arousal, tangage, lacet, roulis, angle d'inclinaison du regard, distance à l'écran] (voir Fig. 1 - gauche). Ensemble, ces caractéristiques capturent l'expression émotionnelle et physique de l'étudiant à chaque seconde. Comme chaque étudiant a consacré un temps différent à chaque question, la longueur (nombre d'images) de la trajectoire varie selon la paire étudiant-question. Pour comparer les trajectoires entre les étudiants, nous avons utilisé la méthode DTW pour les aligner temporellement. Nous avons ainsi calculé les distances par paire entre les trajectoires, ce qui nous indique le degré de similarité de deux SEPT de réponse émotionnelle/physique. Le passage d'un seul point de données moyenné par question à cette

série chronologique à 7 dimensions constitue une avancée méthodologique majeure pour l'analyse du décrochage scolaire. En résumé, le comportement de chaque étudiant lorsqu'il répond à chaque question est désormais représenté sous la forme d'un SEPT : une séquence alignée dans le temps capturant les changements émotionnels et physiques seconde par seconde.

**Visualisation.** Après normalisation des données de chaque dimension dans l'intervalle [-1,1], nous avons visualisé le SEPT afin d'en extraire des informations analytiques. Compte tenu de la nature multidimensionnelle inhérente de ces trajectoires, nous avons sélectionné la valence, l'arousal et la distance face à l'écran comme axes clés pour visualiser le SEPT. Chaque point de la visualisation correspond à une image enregistrée, liée aux expressions faciales et physique de l'étudiant. Nous avons appliqué l'Agglomerative Hierarchical Clustering avec complete linkage et une approche par matrice de distance précalculée en 7D basée sur des mesures de similarité DTW, ce qui nous a permis de regrouper les étudiants présentant une dynamique émotionnelle et une posture comportementale comparables. Dans la figure 2, le code couleur représente les groupes identifiés, capturant des schémas communs d'évolution des émotions et de la position de la tête tout au long de la session de test. Cette visualisation fournit une représentation temporelle des fluctuations émotionnelles des étudiants et de leur influence potentielle sur les performances cognitives. En analysant ces schémas, nous cherchons à identifier des indicateurs comportementaux corrélés aux résultats scolaires.

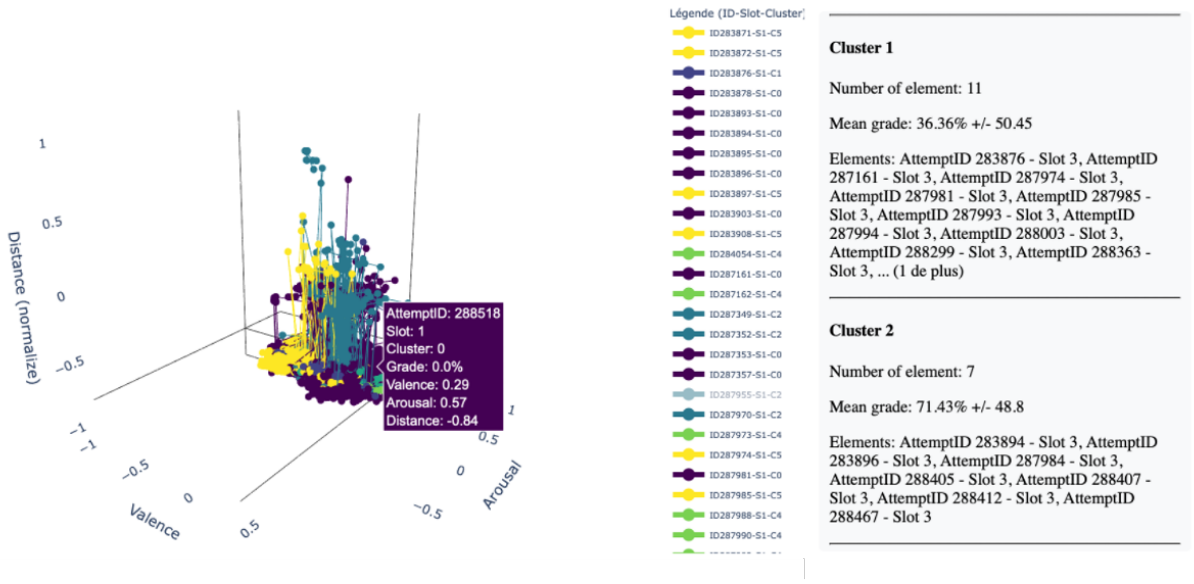



FIGURE 2 – Visualisation 3D du SEPT pour la question 7 du quiz, utilisant la valence, l'arousal et la distance à l'ecran. Les points représentent les données horodatées des étudiants, codées par couleur par l'Agglomerative Hierarchical Clustering dans l'espace DTW 7D. Les résumés des clusters à droite indiquent le nombre d'éléments, la note moyenne et les tentatives associées.

# 4 Résultats et discussion

Afin d'évaluer la pertinence et le pouvoir discriminant du SEPT, nous avons réalisé des analyses de clustering à deux niveaux, parmi les étudiants et entre eux pour chaque quiz. L'objectif était de déterminer si les dynamiques émotionnelles et physiques observées s'expliquaient mieux par des schémas individuels ou par la difficulté des questions, et de valider nos quatre hypothèses (H1–H4).

**H1. La valence et l'arousal seuls sont insuffisants.** Des observations antérieures révèlent une déconnexion entre la valence/l'arousal émotionnel et les résultats de performance, remettant en question la suffisance de ces dimensions en tant qu'indicateurs autonomes. Plusieurs étudiants ayant obtenu de faibles scores ont présenté des trajectoires de valence systématiquement positive ou neutre, et de même, les étudiants performants n'ont pas toujours affiché des niveaux élevés d'arousal ou de valence. Pour valider statistiquement ces résultats, nous avons réalisé une ACP sur des caractéristiques comportementales multimodales et visualisé leurs contributions aux deux principales composantes. En prenant le questionnaire 4231, question 2, comme exemple représentatif, la figure 3 montre que la valence et l'arousal présentent une forte corrélation et des vecteurs relativement courts dans le cercle de corrélation, ce qui signifie qu'ils expliquent moins la variance des données que d'autres caractéristiques comme la rotation de la tête ou la distance à l'écran. Ces dimensions ont peu contribué à la formation de groupes d'étudiants. Cela confirme l'hypothèse H1 : si la valence et l'arousal apportent des informations, elles ne constituent pas à elles seules des prédicteurs fiables de la difficulté perçue. Un ensemble plus large d'indicateurs, incluant la posture physique, doit être pris en compte.

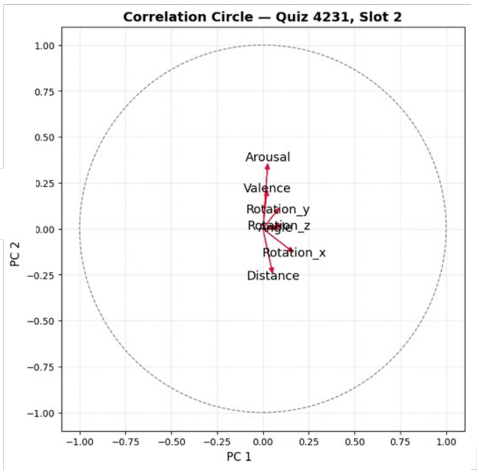



FIGURE 3 – Cercle de corrélation des variables comportementales (espace ACP), Quiz 4231 - question 2. La longueur et la direction de chaque flèche représentent la corrélation d'une caractéristique avec les deux composantes principales.

**H2. Les émotions varient selon les questions.** Pour tester H2, nous avons appliqué un regroupement par question sur SEPT, extrait pour chaque paire (quiz, question). L'objectif était de déterminer si les étudiants ayant des performances similaires, notamment ceux en difficulté, présentaient également des schémas affectifs similaires pour la même question. Sur les 13 questions analysées, 7 ont montré une nette distinction entre les étudiants performants et les étudiants peu performants, avec des profils émotionnels distincts.

Par exemple :

- Le quiz 4233, question 3 (p-value = 0,051) a montré une nette distinction : un groupe d'étudiants obtenant seulement 37,5 % (Cluster 0), un autre 83,3 % (Cluster 1) et un troisième 50 % (Cluster 2). (Fig. 4.) Ce contraste marqué, combiné au regroupement affectif/physique, suggère que les étudiants en difficulté ont réagi avec des schémas émotionnels similaires, reflétant probablement de la frustration ou une déconcentration. De même, la question 4 (p-value = 0,348), bien que moins prononcée, a révélé un groupe d'étudiants obtenant des résultats systématiquement inférieurs (61,8 %), confirmant la présence d'une convergence affective induite par la tâche. Ces résultats renforcent l'hypothèse H2 en confirmant que les émotions varient selon la question et que les étudiants peu performants présentent des trajectoires affectives comparables lorsqu'ils sont mis au défi.
- Quiz 4231, question 5 (p-value = 0,092) : Un groupe d'étudiants peu performants a montré une forte activation. Ces schémas reflètent une mobilisation émotionnelle intense, renforçant l'idée que la difficulté déclenche des réponses physiologiques et affectives communes.
- Quiz 4233, question 1 (p-value = 0,039) a révélé deux grands groupes d'étudiants obtenant des résultats inférieurs à 50 %, et un groupe atypique composé d'un seul étudiant très performant. Bien que ce regroupement mette en évidence une difficulté générale chez la plupart des étudiants, le test statistique ne confirme pas l'existence d'une structure affective significative entre les groupes de performance. Néanmoins, la prévalence de faibles scores peut néanmoins indiquer un stress émotionnel partagé en réponse au contenu de la question.

Dans ces exemples, les réponses émotionnelles ne sont pas idiosyncrasiques : les étudiants en difficulté ont tendance à adopter des comportements convergents face à la même question. Ces résultats valident l'hypothèse H2, montrant que les réponses affectives varient non seulement d'un étudiant à l'autre, mais aussi d'une question à l'autre, avec des schémas cohérents émergeant en réponse à la difficulté perçue ou à la charge cognitive.

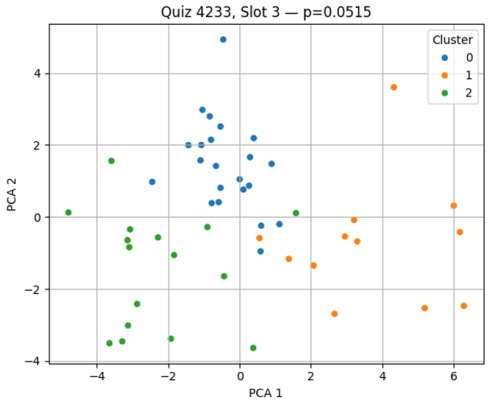



FIGURE 4 – Projection PCA des trajectoires des étudiants pour le questionnaire 4233, question 3. Trois groupes distincts émergent (p-value = 0,0515), révélant des états émotionnels et physiques cohérents parmi les étudiants peu performants, à l'appui de H2.

**H3. Les trajectoires émotionnelles dépendent du contexte et ne sont pas des constantes propres à chaque étudiant.** Les résultats du regroupement par étudiant ont également révélé que les étudiants ne présentaient pas un état affectif unique et récurrent d'une question à l'autre. Au contraire, les

trajectoires étaient façonnées par la complexité spécifique de chaque question. L'état émotionnel et physique d'un étudiant lors d'une question à choix multiples facile différait sensiblement de son état lors d'une tâche de codage complexe, même réalisée au cours de la même séance. Cela contredit la notion de profils affectifs fixes et conforte l'hypothèse H3 : les réponses affectives dépendent du contexte plutôt qu'elles ne sont intrinsèques à l'apprenant.

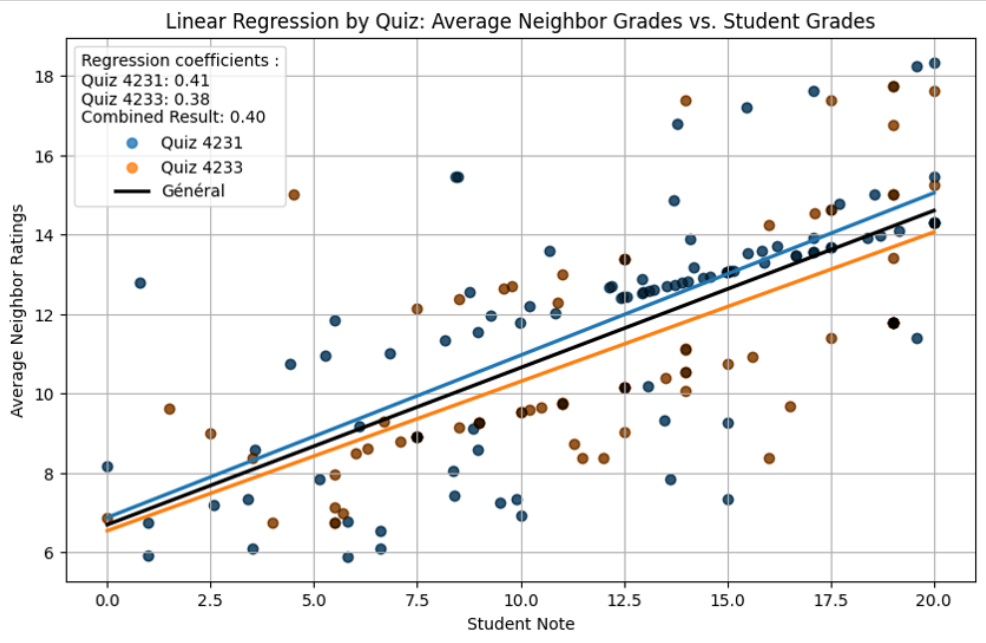



FIGURE 5 – Nuage de points illustrant la corrélation entre les scores individuels des étudiants et les scores moyens de leurs pairs du groupe pour les quiz 4231 et 4233.

**H4. Schémas communs aux questions difficiles.** Afin de tester la convergence interindividuelle face à des défis cognitifs partagés, nous avons appliqué un regroupement par question : le regroupement des trajectoires de tous les étudiants sur une question spécifique. Une structure beaucoup plus claire est apparue. Les groupes composés d'étudiants peu performants présentaient systématiquement des caractéristiques de trajectoire communes, telles qu'une valence décroissante, des mouvements fréquents de la tête et une proximité réduite avec l'écran, tandis que les groupes très performants présentaient des schémas affectifs et physiques plus stables. Quantitativement, ces groupes étaient significativement alignés avec la performance : pour le quiz 4231, la corrélation de Pearson entre la note d'un étudiant et le score moyen de ses pairs était de $r = 0{,}41$ (p-value $< 0{,}001$) ; pour le quiz 4233, $r = 0{,}38$ (p-value $< 0{,}001$). Sur l'ensemble des données du quiz, cette corrélation est restée stable autour de $r \approx 0{,}40$, indiquant que les étudiants d'un même groupe de trajectoire avaient tendance à obtenir des scores similaires. Ces regroupements cohérents et alignés sur les performances (Fig. 5) corroborent l'hypothèse H4 : confrontés à la même question difficile, les étudiants ont tendance à présenter des réactions émotionnelles et physiques similaires, qui se manifestent par des groupes homogènes dans l'espace SEPT. Ainsi, SEPT capture non seulement les réactions individuelles, mais aussi des schémas de difficulté communs, offrant ainsi un moyen évolutif de détecter les contenus problématiques en temps réel.

# 5 Conclusion

Cette étude a présenté le SEPT comme un nouveau cadre pour la capture et l'analyse de données comportementales et affectives fines lors des évaluations en ligne. Grâce à des trajectoires dynamiques et multivariées combinant la valence, l'arousal, l'orientation de la tête et la distance face à l'écran, nous avons démontré que le comportement et la difficulté des étudiants peuvent être déduits au niveau de chaque question. Nos analyses de clustering ont validé les hypothèses clés : (H1) la valence et l'arousal seuls ne suffisaient pas à saisir l'état physique des étudiants, ce qui enrichissait l'interprétation de la confusion ou de l'effort cognitif ; (H2) les réponses émotionnelles et physiques variaient significativement d'une question à l'autre, confirmant que l'émotion est un processus dynamique et contextuel plutôt qu'un phénomène stable à l'échelle de l'évaluation ; (H3) l'absence de signatures individuelles stables au cours d'un quiz a confirmé que ces trajectoires sont davantage façonnées par les exigences spécifiques à la question que par les traits intrinsèques des étudiants ; et (H4) l'émergence de schémas comportementaux communs en réponse aux questions difficiles a révélé que les étudiants en difficulté sur une même question présentaient souvent des profils SEPT similaires, correspondant à des performances faibles et homogènes au sein des groupes. Bien que ces résultats confirment le potentiel discriminant du SEPT, plusieurs limites doivent être reconnues. Premièrement, la fusion de données multimodales émotionnelles et posturales reste techniquement difficile, notamment pour synchroniser et pondérer des entrées hétérogènes. Les travaux futurs devraient explorer les architectures de fusion basées sur l'apprentissage profond, en utilisant des mécanismes d'attention temporelle ou des réseaux neuronaux récurrents, afin d'affiner l'interprétabilité et la robustesse (D'mello & Kory, 2015). Deuxièmement, nos résultats sont actuellement limités à un domaine académique spécifique (apprentissage automatique) et à une population étudiante homogène. Une validation plus large entre les matières, les institutions et les contextes culturels est nécessaire, ainsi que des études longitudinales pour déterminer si les trajectoires du SEPT prédisent les résultats d'apprentissage à long terme (Rodrigo *et al.*, 2012). Pour remédier à ces limites et améliorer le cadre du SEPT, nous identifions quatre axes de recherche futurs. Premièrement, nous cherchons à dériver un indicateur d'engagement en temps réel en transformant le mouvement de la tête, la distance visage-écran et la dynamique faciale en un score d'engagement composite. Des travaux récents utilisant des réseaux convolutifs de graphes spatio-temporels ont montré des résultats prometteurs dans la capture de l'engagement à partir de repères faciaux (Mangaroska *et al.*, 2021). Deuxièmement, nous proposons d'étendre le SEPT pour mesurer la concentration cognitive, en intégrant les changements d'attention via la posture de la tête et le suivi du regard comme indicateurs de concentration ou de distraction pendant la résolution de problèmes (Abedi & Khan, 2024). Troisièmement, pour enrichir la diversité des signaux, le SEPT pourrait intégrer des sources de données multimodales telles que les signaux physiologiques (par exemple, la variabilité du rythme cardiaque) et le comportement numérique (par exemple, la dynamique de la souris ou du clavier), qui se sont avérés améliorer la précision de la reconnaissance des affects (Gaudi *et al.*, 2022). Enfin, des mécanismes de rétroaction en temps réel devraient être développés pour fournir des informations exploitables aux instructeurs ou aux systèmes de tutorat intelligents, en adaptant dynamiquement les stratégies pédagogiques en fonction des trajectoires des étudiants.

# Références


ABEDI A. & KHAN S. S. (2024). Engagement Measurement Based on Facial Landmarks and Spatial-Temporal Graph Convolutional Networks. arXiv :2403.17175 [cs], DOI : 10.48550/arXiv.2403.17175.

ARROYO I., WOOLF B. P., BURELSON W., MULDNER K., RAI D. & TAI M. (2014). A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. *International Journal of Artificial Intelligence in Education*, **24**(4), 387–426. DOI : 10.1007/s40593-014-0023-y.

BAKER R. S. J. D., D'MELLO S. K., RODRIGO M. M. T. & GRAESSER A. C. (2010). Better to be frustrated than bored : The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, **68**(4), 223–241. DOI : 10.1016/j.ijhcs.2009.12.003.

BLIKSTEIN P. & WORSLEY M. (2016). Multimodal Learning Analytics and Education Data Mining : using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, **3**(2), 220–238. Number : 2, DOI : 10.18608/jla.2016.32.11.

BOSCH N. & D'MELLO S. (2017). The affective experience of novice computer programmers. *International Journal of Artificial Intelligence in Education*, **27**(1), 181–206. Place : Germany Publisher : Springer, DOI : 10.1007/s40593-015-0069-5.

BOSCH N., D'MELLO S. K., BAKER R. S., OCUMPAUGH J., SHUTE V., VENTURA M., WANG L. & ZHAO W. (2016). Detecting student emotions in computer-enabled classrooms. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, p. 4125–4129, New York, New York, USA : AAAI Press.

BULAT A., CHENG S., YANG J., GARBETT A., SANCHEZ E. & TZIMIROPOULOS G. (2022). Pre-training strategies and datasets for facial representation learning. arXiv :2103.16554 [cs] version : 2, DOI : 10.48550/arXiv.2103.16554.

CHAUBEY A. & BHATTACHARYA B. (2015). Learning Management System in Higher Education. *IJSTE - International Journal of Science Technology & Engineering* |, **2**, 158–162.

D'MELLO S. K. & KORY J. (2015). A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Comput. Surv.*, **47**(3), 43 :1–43 :36. DOI : 10.1145/2682899.

FRENKEL J., CAJAR A., ENGBERT R. & LAZARIDES R. (2024). Exploring the impact of nonverbal social behavior on learning outcomes in instructional video design. *Scientific Reports*, **14**. DOI : 10.1038/s41598-024-63487-w.

GARRISON D. R., ANDERSON T. & ARCHER W. (1999). Critical Inquiry in a Text-Based Environment : Computer Conferencing in Higher Education. *The Internet and Higher Education*, **2**(2), 87–105. DOI : 10.1016/S1096-7516(00)00016-6.

GAUDI G., KAPRALOS B., COLLINS K. & URIBE QUEVEDO A. (2022). Affective computing : An introduction to the detection, measurement, and current applications. DOI : 10.1007/978-3-030-80571-5_3.

GRAESSER A. C. & D'MELLO S. (2012). Chapter Five - Emotions During the Learning of Difficult Material. In B. H. ROSS, Éd., *Psychology of Learning and Motivation*, volume 57 de *The Psychology of Learning and Motivation*, p. 183–225. Academic Press. DOI : 10.1016/B978-0-12-394293-7.00005-4.

HARLEY J. M., BOUCHET F., HUSSAIN M. S., AZEVEDO R. & CALVO R. (2015). A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*, **48**, 615–625. DOI : 10.1016/j.chb.2015.02.013.

HARLEY J. M., LAJOIE S. P., FRASSON C. & HALL N. C. (2017). Developing Emotion-Aware, Advanced Learning Technologies : A Taxonomy of Approaches and Features. *International Journal of Artificial Intelligence in Education*, **27**(2), 268–297. DOI : 10.1007/s40593-016-0126-8.

IEAMSAARD J., CHAROENSOOK S. N. & YAMMEN S. (2021). Deep Learning-based Face Mask Detection Using YoloV5. In *2021 9th International Electrical Engineering Congress (iEECON)*, p. 428–431. DOI : 10.1109/iEECON51072.2021.9440346.

JIANG Y., BOSCH N., BAKER R. S., PAQUETTE L., OCUMPAUGH J., ANDRES J. M. A. L., MOORE A. L. & BISWAS G. (2018). Expert Feature-Engineering vs. Deep Neural Networks : Which Is Better for Sensor-Free Affect Detection ? In C. PENSTEIN ROSÉ, R. MARTÍNEZ-MALDONADO, H. U. HOPPE, R. LUCKIN, M. MAVRIKIS, K. PORAYSKA-POMSTA, B. MCLAREN & B. DU BOULAY, Éds., *Artificial Intelligence in Education*, p. 198–211, Cham : Springer International Publishing. DOI : 10.1007/978-3-319-93843-1_15.

LEE J. W. (2020). Impact of Proctoring Environments on Student Performance : Online vs Offline Proctored Exams.

LIU M. & KOLLIAS D. (2019). Aff-Wild Database and AffWildNet. arXiv :1910.05318 [cs], DOI : 10.48550/arXiv.1910.05318.

MANGAROSKA K., SHARMA K., GASEVIC D. & GIANNAKOS M. (2021). Exploring students' cognitive and affective states during problem solving through multimodal data : Lessons learned from a programming activity. *Journal of Computer Assisted Learning*, **38**. DOI : 10.1111/jcal.12590.

MOLLAHOSSEINI A., HASANI B. & MAHOOR M. H. (2019). AffectNet : A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, **10**(1), 18–31. arXiv :1708.03985 [cs], DOI : 10.1109/TAFFC.2017.2740923.

MORENO-MARCOS P. M., MUÑOZ-MERINO P. J., MALDONADO-MAHAUAD J., PÉREZ-SANAGUSTÍN M., ALARIO-HOYOS C. & DELGADO KLOOS C. (2020). Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced MOOCs. *Computers & Education*, **145**, 103728. DOI : 10.1016/j.compedu.2019.103728.

NADAUD E., YAACOUB A., HAIDAR S., GRAND B. & PREVOST L. (2024). Emotion Trajectory and Student Performance in Engineering Education : A Preliminary Study. DOI : 10.1007/978-3-031-59465-6_25.

NANDI A., XHAFA F., SUBIRATS L. & FORT S. (2021). Real-Time Emotion Classification Using EEG Data Stream in E-Learning Contexts. *Sensors*, **21**(5), 1589. Number : 5 Publisher : Multidisciplinary Digital Publishing Institute, DOI : 10.3390/s21051589.

PAQUETTE L., DE CARVALHO A. M. & BAKER R. S. (2014). Towards Understanding Expert Coding of Student Disengagement in Online Learning : 36th Annual Meeting of the Cognitive Science Society, CogSci 2014. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014*, p. 1126–1131. Publisher : The Cognitive Science Society.

PERRY, RAYMOND P. R. P. (2014). Control-Value Theory of Achievement Emotions. In *International Handbook of Emotions in Education*. Routledge. Num Pages : 22.

PIOT M., ALABARBE T., GONZALEZ J., LE BAIL C., PREVOST L., BOURDEAU J., BERNARD F. X., BAKER M. & DETIENNE F. (2019). Joint analysis of verbal and nonverbal interactions in collaborative E-learning. In *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, p. 1–5. DOI : 10.1109/ACIIW.2019.8925033.

RODRIGO M., BAKER R., AGAPITO J., NABOS J., REPALAM M., REYES JR S. & SAN PEDRO C. (2012). The Effects of an Interactive Software Agent on Student Affective Dynamics while Using ;an Intelligent Tutoring System. *IEEE Transactions on Affective Computing*, **3**, 224–236. DOI : 10.1109/T-AFFC.2011.41.

SHI G., CHEN S., LI H., TIAN S. & WANG Q. (2024). A study on the impact of COVID-19 class suspension on college students' emotions based on affective computing model. *Applied Mathematics and Nonlinear Sciences*, **9**(1).

SILVA L. C., DE M.B. OLIVEIRA F. C., DE OLIVEIRA A. C. & DE FREITAS A. T. (2014). Introducing the JLoad : A Java Learning Object to Assist the Deaf. In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, p. 579–583. ISSN : 2161-377X, DOI : 10.1109/ICALT.2014.169.

SPITZER M. W. H. & MOELLER K. (2023). Performance increases in mathematics during COVID-19 pandemic distance learning in Austria : Evidence from an intelligent tutoring system for mathematics. *Trends in Neuroscience and Education*, **31**, 100203. DOI : 10.1016/j.tine.2023.100203.

UNION E. (2018). General Data Protection Regulation (GDPR) – Official Legal Text.

WAGNER N., MÄTZLER F., VOSSBERG S. R., SCHNEIDER H., PAVLITSKA S. & ZÖLLNER J. M. (2024). CAGE : Circumplex Affect Guided Expression Inference. arXiv :2404.14975 [cs] version : 1, DOI : 10.48550/arXiv.2404.14975.

WHITEHILL J., SERPELL Z., LIN Y.-C., FOSTER A. & MOVELLAN J. (2014). The Faces of Engagement : Automatic Recognition of Student Engagementfrom Facial Expressions. *Affective Computing, IEEE Transactions on*, **5**, 86–98. DOI : 10.1109/TAFFC.2014.2316163.

ZHAO J. & ITTI L. (2016). shapeDTW : shape Dynamic Time Warping. arXiv :1606.01601 [cs], DOI : 10.48550/arXiv.1606.01601.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Stimuler la Pensée Étudiante avec l'AQG : Vers une Génération Automatique de Questions de Type Étudiant

Abdelbassat Labeche[1, 2] Sébastien Fournier[1]
(1) LIS Laboratory, Aix-Marseille Université, France
(2) École Supérieure d'Informatique (ESI-SBA), Algérie
a.labeche@esi-sba.dz, sebastien.fournier@univ-amu.fr

## RÉSUMÉ

Les systèmes de génération automatique de questions (AQG) sont largement utilisés dans les contextes éducatifs pour évaluer les connaissances. Ces systèmes se concentrent presque exclusivement sur des questions de type enseignant, structurées et factuelles. Cet article propose une approche novatrice, le Student-AQG, qui vise à simuler des questions spontanées qu'un étudiant réel pourrait poser, reflétant ses incompréhensions, sa curiosité ou ses besoins d'approfondissement. En nous appuyant sur les travaux récents en génération de questions autonomes (Mulla & Gharpure, 2023), nous concevons un système modulaire basé sur des LLMs guidés par du prompt engineering, tenant compte du profil cognitif de l'apprenant. Nous décrivons une stratégie d'évaluation combinant des métriques automatiques et des annotations humaines sur la fluidité, la pertinence et la valeur pédagogique. Ce travail vise à aider les élèves à formuler des questions, développant ainsi leur pensée critique, une compétence essentielle souvent négligée à cause du faible questionnement spontané observé en classe (Chin & Osborne, 2008; Raj *et al.*, 2022).

## ABSTRACT

**Stimulating Student Thinking with AQG : Towards Automatic Generation of Student-Like Questions**

Automatic question generation (AQG) systems are widely used in educational settings to support knowledge assessment. Most existing systems focus almost exclusively on teacher-like questions that are structured and factual. This paper proposes a novel approach, Student-AQG, which aims to simulate spontaneous questions a real student might ask—revealing confusion, curiosity, or the need for clarification. Building on recent research in question generation (Mulla & Gharpure, 2023), we design a modular system that leverages large language models guided by prompt engineering, while adapting to the learner's cognitive profile. We also describe an evaluation framework combining automatic metrics and human annotations, focusing on fluency, relevance, and pedagogical value. This research seeks to assist students in formulating their own questions, enhancing their critical thinking—a key competency often overlooked due to the lack of spontaneous questioning in classrooms (Chin & Osborne, 2008; Raj *et al.*, 2022).

MOTS-CLÉS : Génération automatique de questions, Modèles de langage, Apprentissage actif, Prompt engineering, Pensée critique.

KEYWORDS: Automatic question generation, Language models, Active learning, Prompt engineering, Critical thinking.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 126–132.

# 1 Introduction

Ces dernières années, les grands modèles de langage (LLMs) comme ChatGPT, LLaMA ou Mistral ont transformé l'interaction entre intelligence artificielle et apprentissage humain. Pourtant, en contexte scolaire, les étudiants restent souvent passifs et posent rarement des questions en classe, par manque de confiance ou par crainte du jugement. Il devient donc essentiel de les accompagner dans la formulation de questions, afin de les rendre progressivement autonomes dans cette pratique réflexive.

Un usage prometteur des LLMs en éducation est la génération automatique de questions (AQG) à partir d'un texte source. Les travaux de (Mulla & Gharpure, 2023) montrent que la plupart des systèmes AQG se limitent à des questions factuelles, négligeant les aspects métacognitifs. Or, comme l'ont établi (Ikuta & Maruno, 2004), les étudiants qui posent des questions montrent une compréhension plus profonde des concepts. Notre projet s'inscrit dans cette perspective, en proposant un système original centré sur l'élève. Ce **Student-AQG system** vise à simuler, à l'aide d'un LLM, la capacité d'un étudiant réel à poser des questions ouvertes, imparfaites, mais révélatrices de sa curiosité ou de ses doutes.

L'objectif est de stimuler l'engagement cognitif et d'encourager la pensée critique par la formulation de questions pertinentes, adaptées à leur âge et à leur niveau. Contrairement aux approches classiques centrées sur l'évaluation, notre système cherche à reproduire la dynamique naturelle d'un élève face à un contenu éducatif, et à l'accompagner vers une autonomie dans la création de ses propres questions.

# 2 Problématique

La génération automatique de questions est un domaine actif en traitement automatique des langues (TAL), mais la majorité des approches restent centrées sur l'enseignant (Bulathwela *et al.*, 2023). Si ces méthodes sont efficaces pour créer des QCM ou des quiz, elles n'encouragent pas toujours l'engagement cognitif ni la pensée critique des apprenants.

À l'inverse, un système Student-AQG devrait produire des questions traduisant une curiosité réelle, une compréhension partielle ou un besoin d'approfondissement (Chin & Osborne, 2008). Ces questions ne servent pas seulement à évaluer, mais initient un dialogue avec l'enseignant, stimulant un apprentissage plus interactif et réflexif. Le tableau 1 illustre les différences clés entre les deux approches.

TABLE 1 – Comparaison Teacher-AQG vs Student-AQG

| Élément | Teacher-AQG | Student-AQG |
|---|---|---|
| Finalité | Évaluer la compréhension | Simuler le questionnement naturel |
| Style | Structuré, factuel | Curieux, confus, exploratoire |
| Utilisation | Enseignants, quiz | Élèves, auto-apprentissage |
| Données | SQuAD, RACE | Forums, prompts simulés |

Cependant, guider un LLM à adopter une posture « étudiante » reste un défi majeur. Il faut concevoir des prompts générant des formulations naturelles mais pertinentes (Zamfirescu-Pereira *et al.*, 2023), éviter les questions trop simples ou complexes, évaluer leur valeur pédagogique, et limiter les biais cognitifs ou culturels.

# 3 État de l'art

Cette section présente brièvement les principales approches de génération automatique de questions, en s'appuyant sur les méthodologies recensées par (Mulla & Gharpure, 2023), des systèmes à base de règles aux modèles neuronaux actuels (Lopez *et al.*, 2021).

## 3.1 Approches fondées sur des règles et méthodes neuronales supervisées

Les premières méthodes utilisent des structures syntaxiques et des patrons linguistiques pour transformer des phrases en questions. Elles exploitent des arbres syntaxiques, des patrons sémantiques ou des règles heuristiques issues de corpus, mais manquent de souplesse et de généralisation hors domaine.

Avec des corpus comme SQuAD ou WikiAnswers, des architectures Seq2Seq avec attention ont permis de générer automatiquement des questions à partir de textes et réponses. Des modèles comme T5, BART ou des réseaux LSTM ont été entraînés pour produire des questions proches de celles humaines, évaluées par des métriques telles que BLEU, METEOR ou ROUGE.

## 3.2 Transformers et prompting sans supervision

Des modèles récents tels que GPT-3, LLaMA ou BERT, utilisés en zero-shot ou few-shot via le prompt engineering, génèrent des questions pertinentes sans entraînement dédié. Ces approches s'appuient sur la contextualisation des prompts, la reformulation par rôle ou le raisonnement en chaîne pour simuler un questionnement plus naturel. S'inscrivant dans cette troisième approche, les recherches récentes incluent des travaux sur la génération de questions critiques visant à stimuler la pensée analytique (Figueras & Agerri, 2025), la génération de questions motivées par la curiosité simulant le questionnement d'un apprenant découvrant un nouveau concept (Javaji & Zhu, 2024), et l'utilisation des techniques de prompting pour générer des questions éducatives, notamment avec le jeu de données EduProbe qui privilégie les questions commençant par "Pourquoi" et "Comment" (Maity *et al.*, 2024).

Par rapport aux méthodes précédentes, ces modèles offrent une grande souplesse, mais posent des défis en termes de variabilité, de contrôle du style et d'évaluation de la pertinence des questions.

# 4 Méthodologie

Notre méthodologie repose sur la conception d'un système capable de simuler le questionnement naturel d'un élève à partir d'un texte pédagogique. Ce système, appelé Student-AQG, mobilise des modèles de langage avancés et des techniques modernes de prompt engineering pour générer des questions pertinentes, révélatrices de la curiosité ou des incompréhensions d'un apprenant.

## 4.1 Vue d'ensemble du système

Inspiré par le cadre multi-agents proposé dans (Sun *et al.*, 2024), notre système combine plusieurs modules où le LLM alterne entre rôles génératifs et évaluatifs. Le pipeline comprend quatre grandes étapes successives : (1) le prétraitement linguistique, (2) l'extraction de concepts clés, (3) la génération

de questions via un LLM guidé par des prompts optimisés (Lemeš, 2024), et enfin (4) le filtrage et l'évaluation des questions générées.

Le système prend en compte le **profil cognitif de l'élève**, notamment son âge et son niveau scolaire, pour ajuster la formulation, la complexité et l'intention pédagogique de la question. Cette personnalisation vise à renforcer l'engagement et la pertinence des interactions.

## 4.2 Architecture technique

L'implémentation repose sur une architecture modulaire HuggingFace. Elle comprend un pipeline linguistique, une couche d'abstraction entre modèles, et des connecteurs d'évaluation. Notre étude compare des modèles de tailles variées pour évaluer le compromis entre qualité et ressources.

TABLE 2 – Modèles envisagés pour le Student-AQG

| Modèle | Taille | Type |
|---|---|---|
| GPT-4 | $\sim$1T | Propriétaire |
| Qwen | 14B | Open source |
| Mistral | 7B | Open source |
| Gemma | 2B | Open source |

## 4.3 Ingénierie des prompts

Le guidage du LLM repose sur une ingénierie fine des prompts, adaptée au profil cognitif ciblé. Comme l'ont montré (Zamfirescu-Pereira *et al.*, 2023), la formulation des instructions joue un rôle clé dans la qualité des réponses. Nous nous appuyons sur plusieurs stratégies, dont (Scaria *et al.*, 2024).

Nous utilisons un gabarit dynamique où l'élève est simulé comme locuteur. Par exemple :

```
Joue le rôle d'un(e) étudiant(e) de niveau[niveau] en[domaine].Tu viens
d'apprendre le concept suivant:[concept],mais certains points restent
flous. Instructions : 1. Identifie un aspect difficile ou intéressant
2. Formuleune question à poser à ton enseignant
3. (Avancé uniquement) Expliqueton raisonnement
```

Trois techniques principales enrichissent la génération : a) **L'exemple guidé (Few-shot)** : le modèle reçoit 2 ou 3 exemples authentiques de questions d'élèves ; b) **Le raisonnement en chaîne (Chain-of-Thought)** : selon (Wei *et al.*, 2022), cette méthode aide le modèle à formuler sa pensée séquentiellement, améliorant cohérence et plausibilité, surtout pour les niveaux avancés ; c) **La reformulation par rôle** : le modèle adopte l'identité d'un étudiant curieux face à un nouveau concept, favorisant une formulation plus naturelle et spontanée.

Ces techniques génèrent des questions authentiques. Pour un cours d'IHM, le système produit : « Quelle est la différence entre le système sensoriel et le système cognitif dans l'interface homme machine ? » — imitant le questionnement spontané étudiant.

## 4.4 Extension : aide à la formulation de questions

En complément de la génération automatique de questions, nous envisageons d'intégrer un module interactif destiné à **aider les étudiants à formuler leurs questions**. Ce module proposera des *indices*, des *pistes de réflexion* ou des *mots-clés extraits du texte* pour guider l'élève dans sa réflexion.

L'objectif est de favoriser une démarche active où l'étudiant participe au processus de questionnement, ce qui renforce l'appropriation des connaissances et développe la pensée critique. Cette extension se positionne comme un outil pédagogique interactif et complémentaire au système Student-AQG.

## 4.5 Réduction des biais dans la génération de questions

Les grands modèles de langage (LLMs) peuvent reproduire ou amplifier des biais présents dans leurs données d'entraînement, notamment sur le genre, l'origine culturelle ou les stéréotypes implicites (Navigli *et al.*, 2023). Dans un contexte éducatif, ces biais peuvent compromettre la neutralité des questions générées, en introduisant des formulations inadéquates ou non inclusives.

Pour limiter ces effets, plusieurs mécanismes sont prévus. D'abord, les prompts seront formulés de manière explicite et neutre, en évitant les termes marquant une identité spécifique (par exemple, "l'apprenant" au lieu de formulations genrées). Ensuite, un filtrage automatique analysera chaque question à l'aide d'un lexique de sensibilité (issu de ressources sur les biais linguistiques), couplé à une mesure de similarité sémantique via des embeddings (Sentence-BERT ou KeyBERT). Enfin, une évaluation humaine sera menée sur un échantillon représentatif : un groupe d'enseignants et d'étudiants annotera les questions selon une grille de neutralité, représentativité et absence de stéréotypes. Cette validation mesurera l'accord inter-annotateurs et permettra d'ajuster progressivement prompts ou filtres.

Cette approche vise à concilier le réalisme cognitif des questions étudiantes avec des garanties minimales d'équité et d'inclusivité.

## 4.6 Stratégie d'évaluation

L'évaluation du système Student-AQG repose sur une combinaison d'indicateurs automatiques et d'annotations humaines, en accord avec les recommandations de (Mulla & Gharpure, 2023). Le tableau 3 présente les principales métriques utilisées.

TABLE 3 – Métriques d'évaluation du Student-AQG

| Critère | Méthode d'évaluation | Échelle |
|---|---|---|
| Fluidité | Score GPT-4 (cohérence grammaticale) | 0–1 |
| Pertinence contextuelle | Similarité sémantique (Sentence-BERT) | 0–3 |
| Réalisme cognitif | Annotation humaine (marqueurs de doute, confusion) | 1–5 |
| Valeur pédagogique | Jugement expert (utilité, discussion induite) | 1–5 |
| Similarité avec questions étudiantes | Similarité d'embeddings | 0–1 |
| Détection de biais | Checklist qualitative | Binaire |

L'expérimentation inclura une validation humaine par annotateurs indépendants, avec mesure du niveau d'accord et améliorations itératives du processus.

### 4.7 Limites des jeux de données existants

Les systèmes traditionnels de génération de questions éducatives reposent principalement sur des corpus comme SQuAD ou SciQ (Bulathwela *et al.*, 2023). Bien que ces jeux de données produisent des questions factuelles bien structurées, ils ne reflètent ni la spontanéité ni les marqueurs d'incertitude du questionnement étudiant réel. Leurs questions supposent une compréhension totale du texte et visent l'évaluation, plutôt que l'expression de doutes ou de curiosité.

Comme le montre (Bulathwela *et al.*, 2023), même les modèles pré-entraînés sur des textes scientifiques (S2ORC) reproduisent ces limites. Notre approche Student-AQG contourne ces biais en misant sur le prompt engineering plutôt qu'un apprentissage supervisé, générant ainsi des questions plus authentiques et pédagogiquement pertinentes.

## 5 Considérations éthiques

Notre approche intègre plusieurs préoccupations éthiques essentielles à l'usage des LLMs en contexte éducatif. D'abord, la **protection des données** est assurée : aucune information personnelle ni contenu original n'est conservé. Ensuite, des mécanismes de **filtrage post-génération** sont mis en place pour détecter et limiter les biais culturels ou stéréotypes, comme mentionné précédemment. Enfin, la **transparence** est garantie : les utilisateurs sont informés que les questions proviennent d'une simulation, évitant ainsi toute confusion pédagogique.

## 6 Conclusion et perspectives

Le projet Student-AQG propose une approche innovante pour encourager l'apprentissage actif, en simulant le questionnement spontané d'élèves via des modèles de langage. Contrairement aux systèmes centrés sur l'enseignant, notre démarche cherche à reproduire la diversité, l'imprécision et la curiosité propres aux véritables questions étudiantes. Les travaux de (Elkins *et al.*, 2023) montrent que les LLMs, guidés par des taxonomies comme celle de Bloom, peuvent générer des questions pertinentes. Notre contribution se distingue par son attention au réalisme cognitif de l'élève simulé.

Les perspectives futures incluent l'implémentation complète du pipeline basé sur HuggingFace pour valider tous les composants. Nous visons aussi la constitution d'un jeu de données annoté de vraies questions étudiantes, inspiré de (Ikuta & Maruno, 2004), afin de mieux calibrer nos outils d'évaluation. Enfin, une évaluation à différents niveaux scolaires permettra de mesurer l'adaptabilité du système aux profils d'apprenants et la pertinence pédagogique des questions générées.

## Références

BULATHWELA S., MUSE H. & YILMAZ E. (2023). Scalable educational question generation with pre-trained language models. In N. WANG, G. REBOLLEDO-MENDEZ, N. MATSUDA, O. C. SANTOS & V. DIMITROVA, Éds., *Artificial Intelligence in Education*, p. 327–339, Cham : Springer Nature Switzerland.

CHIN C. & OSBORNE J. (2008). Students' questions : a potential resource for teaching and learning science. *Studies in Science Education*, **44**(1), 1–39. DOI : 10.1080/03057260701828101.

ELKINS S., KOCHMAR E., SERBAN I. & CHEUNG J. C. K. (2023). How useful are educational questions generated by large language models ? In N. WANG, G. REBOLLEDO-MENDEZ, V. DIMITROVA, N. MATSUDA & O. C. SANTOS, Éds., *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, p. 536–542, Cham : Springer Nature Switzerland.

FIGUERAS B. C. & AGERRI R. (2025). Benchmarking critical questions generation : A challenging reasoning task for large language models. https://synthical.com/article/057ecc00-f754-4396-b66b-a22f7393884e.

IKUTA J. & MARUNO S. (2004). Do elementary school students come up with questions during class ? *Graduate School of Human-Environment Studies, Kyushu University*. DOI : 10.15017/3566.

JAVAJI S. R. & ZHU Z. (2024). What would you ask when you first saw $a^2 + b^2 = c^2$ ? evaluating llm on curiosity-driven questioning.

LEMEŠ S. (2024). Prompt engineering. In *Artificial Intelligence in Industry 4.0 : The Future That Comes True*, p. 159–170. DOI : 10.5644/PI2024.215.08.

LOPEZ L. E., CRUZ D. K., CRUZ J. C. B. & CHENG C. (2021). Simplifying paragraph-level question generation via transformer language models. In D. N. PHAM, T. THEERAMUNKONG, G. GOVERNATORI & F. LIU, Éds., *PRICAI 2021 : Trends in Artificial Intelligence*, p. 323–334, Cham : Springer International Publishing.

MAITY S., DEROY A. & SARKAR S. (2024). Harnessing the power of prompt-based techniques for generating school-level questions using large language models. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, p. 30–39, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3632754.3632755.

MULLA N. & GHARPURE P. (2023). Automatic question generation : a review of methodologies, datasets, evaluation metrics, and applications. *Progress in Artificial Intelligence*, **12**(1), 1–32. DOI : 10.1007/s13748-023-00295-9.

NAVIGLI R., CONIA S. & ROSS B. (2023). Biases in large language models : Origins, inventory, and discussion. *J. Data and Information Quality*, **15**(2). DOI : 10.1145/3597307.

RAJ T., CHAUHAN P., MEHROTRA R. & SHARMA M. (2022). Importance of critical thinking in the education. *World Journal of English Language*, **12**(3), 126–135. DOI : 10.5430/wjel.v12n3p126.

SCARIA N., DHARANI CHENNA S. & SUBRAMANI D. (2024). Automated educational question generation at different bloom's skill levels using large language models : Strategies and evaluation. In A. M. OLNEY, I.-A. CHOUNTA, Z. LIU, O. C. SANTOS & I. I. BITTENCOURT, Éds., *Artificial Intelligence in Education*, p. 165–179, Cham : Springer Nature Switzerland.

SUN H., LIU Y., WU C., YAN H., TAI C., GAO X., SHANG S. & YAN R. (2024). Harnessing multi-role capabilities of large language models for open-domain question answering. In *Proceedings of the ACM Web Conference 2024*, p. 4372–4382.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E., LE Q. V. & ZHOU D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO & A. OH, Éds., *Advances in Neural Information Processing Systems*, volume 35, p. 24824–24837 : Curran Associates, Inc.

ZAMFIRESCU-PEREIRA J., WONG R. Y., HARTMANN B. & YANG Q. (2023). Why johnny can't prompt : How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, p. 1–21. DOI : 10.1145/3544548.3581388.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Un outil conversationnel basé sur un graphe de connaissances, des LLM et un modèle BERT pour les programmes d'alternance en France

Baba Mbaye[1] Diana Nurbakova[2] Duaa Baig[1, 2]
(1) Effet B, 1 Rue Dr Fleury Papillon, 69100 Villeurbanne, France
(2) INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 20 Avenue Albert Einstein, 69621 Villeurbanne, France
baba@effetb.com, diana.nurbakova@insa-lyon.fr, duaa.baig@insa-lyon.fr

## RÉSUMÉ

Le suivi efficace de l'acquisition des compétences dans les programmes de l'alternance, présente des défis importants pour la technologie éducative. Cet article présente un nouvel agent conversationnel intégré dans un livret de formation numérique qui relève ces défis grâce à une architecture multimodale. Notre système intègre (1) un graphe de connaissances spécifique à un domaine, lié à des référentiels de compétences, (2) des grands modèles de langage (LLM) et (3) un composant génératif basé sur BERT. Cette approche hybride permet à la fois une représentation structurée des trajectoires d'apprentissage et des capacités d'interaction en langage naturel, ce qui permet un suivi nuancé des progrès et des interventions personnalisées. L'évaluation empirique démontre que le système fournit un retour d'information contextuellement pertinent qui s'adapte aux modèles d'apprentissage individuels, ce qui permet une acquisition plus efficace des compétences.

## ABSTRACT

**A Conversational Tool Based on Knowledge Graph, LLMs and BERT Model for Work-Study Programs in France**

Effective skill acquisition monitoring in vocational training programs, particularly apprenticeships, presents significant challenges for educational technology. This paper introduces a novel conversational agent embedded within a digital training booklet that addresses these challenges through a multimodal architecture. Our system integrates (1) a domain-specific knowledge graph linked to competency frameworks, (2) state-of-the-art large language models (LLMs) and (3) a BERT-based generative component. This hybrid approach enables both structured representation of learning trajectories and natural language interaction capabilities, allowing for nuanced progress monitoring and personalized interventions. Empirical evaluation demonstrates that the system provides contextually relevant feedback that adapts to individual learning patterns, resulting in more efficient skill acquisition.

MOTS-CLÉS : Outil conversationnel, Graphe de connaissances, LLMs, BERT, programmes d'alternance.

KEYWORDS: Conversational tool, Knowledge graph, LLMs, BERT, Work-Study Programs.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 133–144.

# 1 Introduction

Les programmes en alternance combinent l'apprentissage académique et l'expérience professionnelle. Ils offrent de nombreux avantages, comme l'acquisition de compétences concrètes liées au domaine d'études, l'amélioration de l'employabilité des étudiants, et la possibilité d'appliquer des concepts théoriques dans des situations réelles (University of Notre Dame, Australia *et al.*, 2019). Cependant, ces programmes rencontrent également plusieurs défis : la gestion du temps, la taille des promotions (University of Notre Dame, Australia *et al.*, 2019), les difficultés de placement en entreprise (Stanley & Xu, 2019), ou encore l'équilibre entre le travail et les études (Uclaray *et al.*, 2023).

Chaque programme de formation couvre un ensemble de compétences et de savoirs propres au métier visé. Traditionnellement, un livret de suivi est utilisé pour suivre les compétences acquises par les apprenants. Ce livret est consulté par les apprentis eux-mêmes, les tuteurs en entreprise, les formateurs de l'établissement scolaire et les responsables pédagogiques. Avec la transition numérique, ces livrets sont désormais disponibles sous forme digitale, permettant d'évaluer les progrès des apprenants et de valider les compétences. Cependant, ces versions numériques restent souvent statiques et nécessitent une saisie manuelle régulière. Pour simplifier les interactions avec le livret et limiter ces interventions manuelles, l'intégration d'un outil conversationnel intelligent peut être une solution pertinente (Pérez *et al.*, 2020; Wollny *et al.*, 2021).

Les chatbots sont de plus en plus utilisés dans le domaine de l'éducation pour enrichir l'expérience d'apprentissage et apporter un soutien personnalisé. Ils peuvent servir d'assistants pédagogiques, aider à évaluer les étudiants et fournir un retour d'information en temps réel (Pérez *et al.*, 2020; Georgescu, 2018; Chamorro-Atalaya *et al.*, 2023). De plus, ils permettent un accompagnement individualisé en s'adaptant aux besoins spécifiques de chaque apprenant (Wollny *et al.*, 2021; Alfehaid & Hammami, 2023; Ramandanis & Xinogalos, 2023). Pourtant, ces outils conversationnels n'ont pas encore été intégrés aux livrets numériques dans le cadre des programmes en alternance.

Pour combler ce manque, nous proposons un agent intelligent basé sur trois éléments principaux :

1. Un graphe de connaissances (GC) pour modéliser les compétences des apprentis et les programmes en alternance.
2. Un modèle de langage de grande taille (Large Language Model - LLM) pour analyser les requêtes des utilisateurs et les traduire en requêtes Cypher afin d'extraire les informations du graphe de connaissances.
3. Un modèle BERT ajusté (fine-tuned) (Devlin *et al.*, 2019) pour générer des réponses précises et adaptées.

Notre approche se concentre sur trois scénarios :
— Sc1, Suivi des compétences acquises ;
— Sc2, Recommandation des compétences à développer ;
— Sc3, Suivi personnalisé et retour d'information.
Le reste du papier est organisé de la façon suivante. Nous présentons le contexte du travail dans la Section 2. Les Sections 3 et 4 décrivent notre outil conversationnel. Nous présentons nos expérimentations dans la Section 5, suivi de la discussion sur les défis et perspectives du travail (Section 6). Enfin, nous concluons le papier avec le résumé des contributions et les travaux futurs (Section 7).

# 2 Organisation et Environnement Éducatif des Formations en Alternance

Depuis les années 1990, la formation en alternance connaît une croissance importante (Le Thuaut, 2005). En 2022, on comptait environ 837 000 apprentis (Ministère du Travail, 2024). Cette augmentation est en grande partie due à la réforme de 2018, qui a renforcé les incitations financières et simplifié les démarches administratives (Assemblée-Nationale, 2018).

## 2.1 Un cadre réglementaire structurant

L'organisation des formations professionnelles est encadrée par *France Compétences* – un organisme public chargé de réguler la qualité et le financement des formations. Cette institution gère également le Répertoire National des Certifications Professionnelles (RNCP), un registre officiel qui recense toutes les qualifications reconnues par l'État. Chaque certification inscrite au RNCP est classée selon un référentiel national en 8 niveaux, correspondant aux différents degrés de qualification (du niveau CAP au doctorat). Ces certifications sont détaillées en blocs de compétences, qui définissent les savoir-faire et les critères d'évaluation attendus.

Un exemple concret est le diplôme d'ingénieur en informatique délivré par l'INSA Lyon, répertorié sous le code RNCP35971. Ce diplôme est structuré en 5 blocs de compétences, parmi lesquels on trouve, par exemple, le bloc RNCP35971BC01, intitulé : "*Concevoir, développer et maintenir des logiciels de qualité*". Chaque bloc précise les compétences à maîtriser, comme la conception d'architectures logicielles complexes ou l'utilisation d'outils de modélisation.

Pour assurer le suivi des apprentissages, des carnets de suivi numériques sont utilisés. Parmi les plateformes les plus courantes, on peut citer Campus skill, GowizApp , Kwark, Sheldon skills et STUDEA. Ces outils permettent de documenter l'acquisition des compétences en temps réel.

## 2.2 Objectifs pédagogiques de notre outil numérique de suivi

L'intégration de notre solution numérique dans les parcours d'alternance répond à plusieurs objectifs pédagogiques :

1. Suivi automatisé des compétences : Ces outils permettent de suivre en temps réel l'évolution des compétences des apprentis, évitant ainsi les lourdeurs administratives liées au suivi manuel. Par exemple, un tuteur peut interroger le système pour savoir : *"Quelles compétences Jean a-t-il validées dans le module de développement logiciel ?"* et obtenir un résumé précis des progrès réalisés.
2. Recommandations personnalisées : Grâce à l'analyse des parcours d'apprentissage, ces systèmes suggèrent des compétences à prioriser en fonction des besoins de chaque apprenti. Par exemple, une question comme : *"Quelles compétences Marie devrait-elle travailler ensuite ?"* permettra au système de proposer un itinéraire de formation adapté.
3. Analyse comparative des performances : Ces outils offrent aussi la possibilité de comparer les performances d'un apprenti avec celles de ses pairs ou avec les attentes du programme. Par exemple, un tuteur peut demander : *"Comment Thomas progresse-t-il par rapport aux attentes ?"* et obtenir une analyse contextualisée.

## 2.3 Un outil au service de plusieurs publics

Ce type de système bénéficie à l'ensemble des acteurs impliqués dans les parcours d'alternance :

1. *Les apprentis* : Ils peuvent consulter en permanence l'état d'avancement de leurs compétences, identifier leurs points forts et les domaines à améliorer. Grâce aux recommandations personnalisées, ils peuvent aussi adapter leur apprentissage tout en restant en phase avec les objectifs de leur formation.
2. *Les tuteurs et formateurs* : Ces outils simplifient le suivi des apprentis en automatisant la collecte des données et en facilitant l'identification rapide des éventuelles difficultés. Cela permet aux encadrants de se concentrer sur des missions à plus forte valeur ajoutée, comme l'accompagnement personnalisé ou le coaching ciblé.
3. *Les établissements de formation* : À une échelle plus large, les données agrégées permettent aux institutions d'ajuster leurs programmes en fonction des évolutions du marché du travail. Cette approche favorise l'adaptation continue des cursus pour répondre aux nouvelles compétences demandées par les entreprises.

En combinant formation théorique et expérience professionnelle, les programmes en alternance jouent un rôle clé dans l'insertion et la montée en compétences des apprenants. L'usage d'outils numériques de suivi transforme ces parcours en rendant le suivi plus efficace, en personnalisant les recommandations et en facilitant l'évaluation des compétences. Cette approche innovante allège la charge administrative tout en renforçant la qualité de l'accompagnement pédagogique, au bénéfice de tous les acteurs de l'alternance.

# 3 Architecture Technique du Système Conversationnel

Cette section décrit les trois principaux composants de notre outil conversationnel : un graphe de connaissances (GC) modélisant le livret de formation numérique, des modèles de langage étendus (LLMs) pour l'analyse des requêtes utilisateurs, et un modèle BERT ajusté pour la génération de réponses.

## 3.1 Livret de Formation Numérique et Graphe de Connaissances

Le livret de formation numérique est un système dématérialisé de suivi des progrès des apprentis. Il automatise la collecte et l'organisation des données liées à la formation. Nous modélisons ce livret sous la forme d'un graphe de connaissances (GC) (Hogan *et al.*, 2022; Qu *et al.*, 2024) construit à partir de deux sources de données principales :

1. Données de suivi des progrès des apprentis issues de notre système de livret numérique.
2. Données de la plateforme France Compétences sur les compétences, qualifications et unités de formation.

Notre GC est composé d'entités/nœuds interconnectés représentant différents aspects du processus d'apprentissage :

— `Compétences` (5 020 nœuds) : Chaque compétence validée au sein du programme de formation est représentée comme un nœud, contenant des attributs tels que l'identifiant de la compétence, le niveau de compétence (aligné sur le cadre RNCP) et les critères d'évaluation.

— `RNCP` (188 nœuds) : Nœuds représentant les certifications reconnues par le Répertoire National des Certifications Professionnelles (RNCP).
— `Catégories de Formation` (170 nœuds) : Classifications générales regroupant des programmes connexes, comme "Informatique et Développement Logiciel."
— `Cours de Formation` (508 nœuds) : Programmes de formation spécifiques liés aux qualifications du RNCP.
— `Apprentis` (2 691 nœuds) : Chaque apprenti est un nœud avec des attributs suivant leurs progrès, les compétences validées et les dates d'acquisition.
— `Formateurs/Tuteurs` (2 010 nœuds) : Les formateurs des établissements éducatifs et des entreprises sont modélisés sous forme de nœuds liés aux apprentis.
— `Écoles` (46 nœuds) : Institutions offrant des programmes de formation liés aux compétences validées.
— `Unités de Formation en Apprentissage (UFA)` (413 nœuds) : Divisions spécialisées des écoles dédiées aux apprentissages.

Le KG capture les relations entre ces entités à travers des arêtes, notamment :
— `HAS_SKILL` (50 738 arêtes) : Relie les apprentis aux compétences acquises ;
— `ATTENDS` (2 903 arêtes) : Relie les apprentis aux cours de formation suivis ;
— `HAS_RNCP` (226 arêtes) : Relie les écoles aux certifications RNCP correspondantes.

Nous avons implémenté le GC en utilisant Neo4J, ce qui nous permet de suivre les parcours d'acquisition des compétences et de surveiller la progression des apprentis. Par exemple, lorsqu'un utilisateur demande : *"Quelles compétences l'apprenti Jean devrait-il prioriser ensuite ?"*, le système suit les chemins de relation pour identifier les compétences non validées pertinentes pour le programme de Jean.

## 3.2 LLMs pour l'Interprétation des Requêtes et la Reconnaissance d'Intentions

Nous exploitons des modèles de langage étendus (LLMs) comme GPT-4 pour l'interprétation des requêtes et la reconnaissance d'intentions (Wan *et al.*, 2024; Kumar, 2024). Bien que les LLMs excellent dans la compréhension du langage naturel, ils peuvent produire des affirmations factuellement incorrectes (Wang *et al.*, 2024). Pour pallier ce problème, notre système intègre le raisonnement des LLMs avec le GC en utilisant un cadre *Planification-Récupération-Raisonnement inspiré de Luo et al. (Luo* et al.*, 2023)*.

Le traitement des requêtes utilisateur se fait en plusieurs étapes. D'abord, lors de la phase de planification, le modèle de langage large (LLM) identifie l'intention de l'utilisateur et génère des chemins de relation qui serviront de base pour la récupération des informations. Ensuite, durant la phase de récupération, les données pertinentes sont extraites du GC en suivant ces chemins de relation définis précédemment. Une fois les informations nécessaires récupérées, un modèle BERT ajusté intervient lors de la phase de raisonnement pour générer des réponses contextuelles et adaptées à la requête. Enfin, dans la phase de génération de réponse, le LLM produit une réponse finale en langage naturel, fluide et compréhensible pour l'utilisateur, tout en restant fidèle aux données extraites du GC.

Dans notre étude, nous avons évalué sept modèles de langage étendu (LLMs), à savoir GPT-4 (Achiam *et al.*, 2023), LLaMA-2 7B (Touvron *et al.*, 2023), GPT-J-6B (Wang, 2021), Falcon 7B (Almazrouei *et al.*, 2023), BLOOM (Workshop *et al.*, 2022) et Mistral 7B(Jiang, 2024), en fonction de plusieurs critères tels que la disponibilité, la précision et le score F1. Cette évaluation s'est basée

sur trois ensembles de données distincts : le premier, Sc1 : Open SQuAD (https://tinyurl.com/2yup27s8), qui comprend 27 713 requêtes ; le deuxième, Sc2 (https://tinyurl.com/muw7vm96) : Ensemble de Données sur les Carrières ; et le troisième, Sc3 (https://tinyurl.com/bdh23acu) : Ensemble de Données ASSISTments 2012-2013. Ces évaluations nous ont permis de comparer les performances des modèles dans des contextes variés et de choisir celui qui répondait le mieux à nos besoins en termes de précision et d'efficacité.

TABLE 1 – LLM comparaison

| Modèle | GPT-4 | LLaMA-2 7B | GPT-J-6B | Falcon 7B | BLOOM | Mistral 7B |
|---|---|---|---|---|---|---|
| Disponibilité | **Payant** | Gratuit | Gratuit | Gratuit | Gratuit | Gratuit |
| Précision (%) | **84** | 82 | 81 | 78 | 73 | 72 |
| Score F1 (%) | **81** | 80 | 77 | 77 | 71 | 71 |

GPT-4 a obtenu la meilleure précision (84%) et le meilleur score F1 (81%), ce qui en fait le choix optimal pour notre implémentation. Sa capacité à structurer des réponses basées sur des données interconnectées s'aligne parfaitement avec notre GC.

## 3.3 Modèle de Raffinement basé sur BERT

Nous utilisons un modèle BERT (Devlin *et al.*, 2019) ajusté pour la génération de réponses et la reconnaissance d'intentions. Bien que les LLMs excellent dans la compréhension des requêtes, BERT garantit des réponses concises, précises et ancrées dans le GC.

Le modèle BERT est ajusté à l'aide de trois principales sources de données. La première source provient des données anonymisées du Livret de Formation Numérique qui sont issues de notre GC. La deuxième source est la documentation sur la Formation Professionnelle fournie par France Compétences, qui contient des informations essentielles sur les compétences et les parcours de formation. Enfin, la troisième source, EdNet, est une base de données éducatives à grande échelle qui inclut des interactions entre les étudiants et les plateformes d'apprentissage (Choi *et al.*, 2020).

Ce processus d'ajustement vise à améliorer les performances de BERT dans deux domaines clés, en affinant ses capacités de compréhension et de génération de réponses.

D'une part, il optimise la génération de réponses, permettant au modèle de produire des réponses plus précises, contextualisées et adaptées aux besoins des utilisateurs. Grâce à un entraînement sur les données du GC, BERT devient plus apte à identifier les informations pertinentes et à formuler des réponses cohérentes, en tenant compte du contexte et des nuances du langage. Cette optimisation contribue à une meilleure expérience utilisateur en réduisant les ambiguïtés et en améliorant la fluidité des échanges.

D'autre part, le processus améliore significativement la reconnaissance des intentions en permettant au modèle de classifier les requêtes avec une grande précision. Plus précisément, BERT est capable d'identifier et de catégoriser les intentions des utilisateurs dans les scénarios Sc1 à Sc3 avec une précision de 92 %) sur notre ensemble de test 1. Cette amélioration est cruciale pour affiner la compréhension du langage naturel et assurer une meilleure adéquation entre la demande de l'utilisateur et la réponse fournie.

Grâce à cet affinage, BERT devient plus performant et mieux adapté aux spécificités du domaine de la formation professionnelle. Il offre ainsi des interactions plus fluides, une meilleure compréhension des requêtes et une capacité accrue à répondre avec pertinence, contribuant à une optimisation globale de son utilisation dans ce contexte particulier.

# 4 Méthodologie

Notre approche repose sur l'intégration de trois composants complémentaires : notre GC pour la représentation structurée des données, GPT-4 pour la compréhension du langage naturel, et un modèle BERT affiné pour la génération des réponses. Cela permet une surveillance intelligente et un accompagnement personnalisé des apprentis, tout en garantissant la précision des informations et la pertinence contextuelle. Notre outil suit une architecture modulaire en pipeline (voir Figure 1), séparant l'analyse de la requête, la récupération d'informations et la génération des réponses. Cette conception modulaire assure une flexibilité accrue et permet une adaptation rapide aux évolutions technologiques. Grâce à cette approche, nous pouvons facilement intégrer de nouvelles fonctionnalités et affiner les modèles en fonction des besoins spécifiques des utilisateurs. De plus, l'architecture en pipeline favorise une meilleure gestion des ressources, optimisant ainsi les performances du système. Cela se traduit par une plus grande réactivité et une amélioration continue de la qualité des réponses fournies. Enfin, cette structure facilite l'évolutivité du système, permettant son déploiement à grande échelle sans compromettre son efficacité et sa précision.

1. **Analyse de la requête (GPT-4)** : lorsqu'un utilisateur soumet une requête, GPT-4 s'occupe d'extraire les entités clés (comme les noms des apprentis, les références de compétences), de déterminer l'intention et d'identifier les paramètres contextuels. Grâce aux capacités de compréhension contextuelle de GPT-4, nous obtenons un taux de précision de 84% dans l'interprétation des intentions des utilisateurs. Cette étape permet de comprendre clairement ce que l'utilisateur recherche, ce qui permet de fournir une réponse plus précise et pertinente.
2. **Raisonnement sur le GC** : la requête traitée est ensuite transformée en une requête Cypher pour récupérer des données structurées à partir du Graph de Connaissances. Cela permet de tirer parti de la structure du graphe pour effectuer des raisonnements complexes, tels que l'analyse des dépendances entre compétences, l'identification des lacunes, la proposition de recommandations personnalisées et des raisonnements basés sur les cohortes. Ce processus garantit que les réponses sont basées sur des données fiables et adaptées à chaque apprenti.
3. **Génération de la réponse (BERT)** : le modèle BERT, préalablement affiné, prend les données structurées extraites du GC et les transforme en une réponse en langage naturel. Ce processus permet de fournir des réponses cohérentes, faciles à comprendre et adaptées au contexte de l'utilisateur. Ainsi, l'outil offre des réponses qui sont à la fois précises et compréhensibles, tout en restant factuellement correctes.
4. **Intégration des retours** : les interactions des utilisateurs sont constamment surveillées pour améliorer la performance du système. Cette intégration des retours permet à l'outil de s'adapter et de se perfectionner au fur et à mesure des échanges, garantissant ainsi une amélioration continue de la qualité des réponses. De plus, un système d'apprentissage automatique permet d'ajuster dynamiquement les modèles en fonction des tendances émergentes et des besoins spécifiques des utilisateurs.
5. **Sécurité et éthique** : un volet essentiel de notre approche repose sur la sécurité et l'éthique dans l'utilisation des données. Nous mettons en place des protocoles stricts de protection des

données personnelles et assurons une conformité avec les réglementations en vigueur. Cela permet de garantir un usage responsable et transparent de l'intelligence artificielle, tout en préservant la confidentialité et l'intégrité des informations traitées.

Ainsi, notre approche modulaire et évolutive permet d'allier performance, adaptabilité et rigueur dans l'accompagnement des apprentis, tout en offrant des solutions adaptées aux défis de l'apprentissage et de la formation continue.

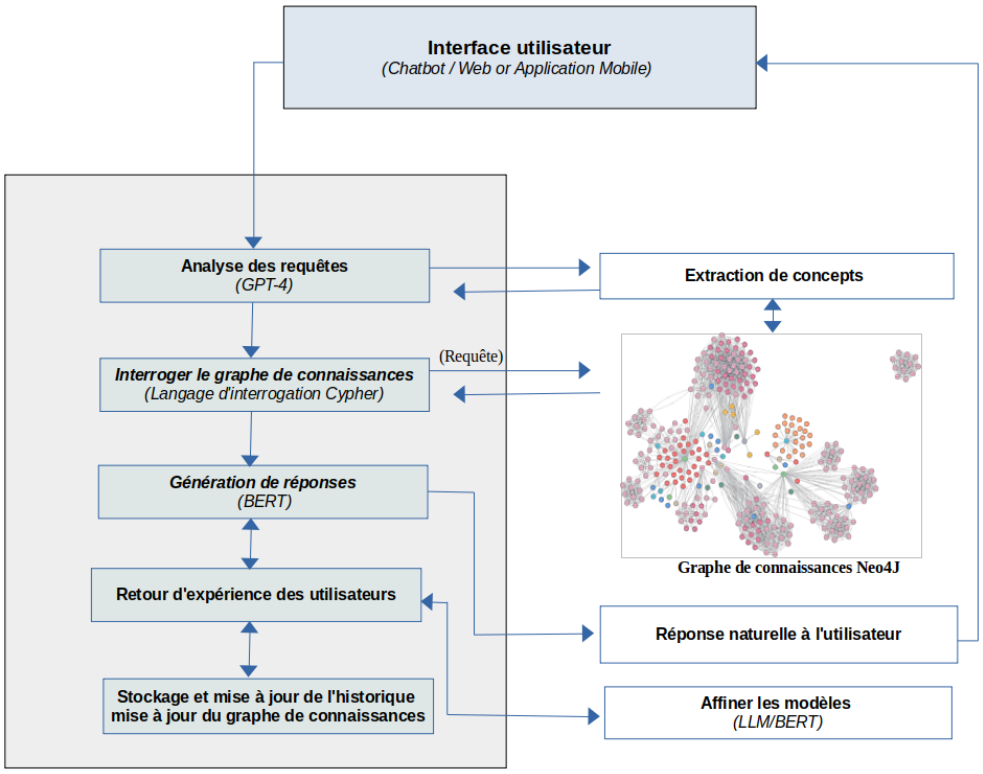



FIGURE 1 – Architecture système

# 5 Expérimentations et résultats

L'objectif de cette expérience était d'évaluer la performance technique d'un outil de conversation. Pour cela, un jeu de données de test composé de 1 500 requêtes a été élaboré, couvrant trois scénarios différents : surveillance des compétences, recommandation de compétences et suivi personnalisé. Les requêtes ont été collectées à partir de trois sources : Open SQuAD (https://tinyurl.com/2yup27s8), le Career Dataset (https://tinyurl.com/muw7vm96) et l'ASSISTments Data Set (https://tinyurl.com/bdh23acu). Quatre configurations ont été comparées pour déterminer l'approche optimale de performance :

1. LLM uniquement : Utilisation de GPT-4 pour comprendre les requêtes et générer des réponses.
2. BERT uniquement : Utilisation de BERT pour comprendre les requêtes et générer des réponses.
3. LLM+BERT : Utilisation de GPT-4 pour comprendre les requêtes et de BERT pour générer les réponses.

4. LLM+BERT+GC : Approche intégrée, combinant GPT-4, BERT et raisonnement basé sur notre GC.

Le tableau 2 ci-dessous montre que les performances des systèmes ont été évaluées en termes de précision et de score F1. L'approche LLM+BERT+GC s'est révélée bien plus performante que les autres configurations sur tous les indicateurs. En particulier, elle a atteint une précision de 93 % dans le scénario de surveillance des compétences.

TABLE 2 – Résultats comparatifs

| Method | LLM-only | BERT-only | LLM+BERT | LLM+BERT+GC |
|---|---|---|---|---|
| Précision (%) | 84 | 83 | 88 | **93** |
| Score F1 (%) | 81 | 80 | 85 | **92** |

L'approche intégrée a montré une grande efficacité dans les situations nécessitant un raisonnement complexe, telles que le recoupement des dépendances entre les compétences, les données de progression individuelles et les performances des groupes.

# 6 Défis et perspectives

Notre solution a permis de mettre en lumière plusieurs considérations clés pour l'intégration d'outils conversationnels basés sur l'IA dans la formation professionnelle. Parmi les défis principaux, on retrouve les problèmes de qualité et de cohérence des données, qui impactent fortement la fiabilité des réponses. Il est également nécessaire d'effectuer un ajustement continu des modèles d'IA afin de suivre l'évolution des programmes de formation. Un autre obstacle majeur réside dans les limitations d'interprétabilité des modèles de langage, ce qui nécessite des mécanismes pour expliquer les recommandations faites aux utilisateurs. L'adoption de ces outils par les utilisateurs pourrait être freinée par la résistance à l'IA ou une maîtrise limitée des outils numériques. De plus, la question de l'atténuation des biais reste primordiale pour garantir des évaluations de compétences justes et équitables.

Malgré ces défis, l'intégration d'outils basés sur l'IA offre des opportunités significatives. Grâce à l'utilisation des modèles de LLM et des GC, notre solution permet de proposer des recommandations personnalisées pour le développement des compétences ainsi qu'un feedback adapté. De plus, elle offre aux formateurs la possibilité d'ajuster leurs stratégies en temps réel en se basant sur des informations précises. Le système automatise le suivi des progrès et la génération de retours, ce qui permet une mise en œuvre à grande échelle et réduit la charge administrative. L'intégration avec des carnets de formation numériques permet de suivre les compétences en temps réel et de suggérer des parcours d'apprentissage optimaux, tout en facilitant la communication entre les différentes parties prenantes. Cette approche hybride de l'IA a un potentiel d'extension au-delà de la formation professionnelle, en s'appliquant également à l'enseignement supérieur, à la formation en entreprise et aux programmes de certification.

# 7 Travaux futurs et conclusion

L'intégration des graphes de connaissances, des Modèles de LLMs et de la génération de réponses basée sur BERT représente une avancée significative dans le soutien aux programmes d'alternance en France. Cette approche combine efficacement des connaissances structurées avec des capacités de traitement du langage naturel, ce qui permet d'obtenir une précision impressionnante de 93%, bien supérieure à celle des composants individuels.

Cependant, malgré ces résultats prometteurs, plusieurs défis demeurent. Tout d'abord, il y a la normalisation des données, un enjeu essentiel pour assurer la cohérence et l'interopérabilité des différentes sources de données. Ensuite, la question de l'adoption par les utilisateurs se pose, car il est crucial d'encourager les étudiants, formateurs et administrateurs à adopter cette technologie, malgré leur familiarité potentiellement limitée avec les outils basés sur l'intelligence artificielle. Enfin, la compréhensibilité du modèle reste un défi important, car il est nécessaire de garantir la transparence des décisions et recommandations produites par le système.

Les opportunités offertes par cette approche sont nombreuses, notamment en matière de parcours d'apprentissage personnalisés. Cela permettrait d'adapter les formations en fonction des besoins et des progrès de chaque apprenant. De plus, cette technologie pourrait grandement améliorer l'efficacité administrative, en facilitant le suivi et l'évaluation des compétences tout au long du programme de formation.

Pour les travaux futurs, nous prévoyons de mener une étude qualitative dans les centres de formation professionnelle afin de recueillir les retours des parties prenantes, d'évaluer leur satisfaction et de mesurer l'impact du système sur les résultats d'apprentissage. Nous souhaitons également intégrer des techniques d'IA explicable (XAI), afin de rendre le système plus transparent en offrant des visualisations du progrès des compétences et des justifications des recommandations proposées. Par ailleurs, nous prévoyons d'étendre les fonctionnalités de notre outil en y intégrant la reconnaissance vocale et en mettant en place un apprentissage par renforcement, permettant d'ajuster dynamiquement les parcours d'apprentissage en fonction des performances des apprenants.

À mesure que les programmes d'alternance gagnent en popularité à l'échelle mondiale, les outils alimentés par l'intelligence artificielle, permettant de personnaliser et de suivre les parcours d'apprentissage, deviendront de plus en plus précieux. Notre travail contribue au domaine de l'IA appliquée à l'éducation, en proposant une mise en œuvre concrète répondant aux défis spécifiques de la formation professionnelle, avec des applications potentielles dans l'enseignement supérieur, la formation en entreprise et les programmes de certification.

# Références

ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S., ANADKAT S. *et al.* (2023). Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*.

ALFEHAID A. & HAMMAMI M. A. (2023). Artificial Intelligence in Education : Literature Review on The Role of Conversational Agents in Improving Learning Experience. *International Journal of Membrane Science and Technology*, **10**(3), 3121–3129. DOI : 10.15379/ijmst.v10i3.3045.

ALMAZROUEI E., ALOBEIDLI H., ALSHAMSI A., CAPPELLI A., COJOCARU R., DEBBAH M., GOFFINET É., HESSLOW D., LAUNAY J., MALARTIC Q. *et al.* (2023). The falcon series of open language models. *arXiv preprint arXiv :2311.16867*.

ASSEMBLÉE-NATIONALE (2018). Etude d'impact sur le texte, n° 904.

CHAMORRO-ATALAYA O., HUARCAYA-GODOY M., DURÁN-HERRERA V., NIEVES-BARRETO C., SUAREZ-BAZALAR R., CRUZ-TELADA Y., ALARCÓN-ANCO R., HUAYHUA-MAMANI H., VARGAS-DIAZ A. & BALAREZO-MARES D. (2023). Application of the Chatbot in University Education : A Systematic Review on the Acceptance and Impact on Learning. *International Journal of Learning, Teaching and Educational Research*, **22**(9), 156–178. DOI : 10.26803/ijlter.22.9.9.

CHOI Y., LEE Y., SHIN D., CHO J., PARK S., LEE S., BAEK J., BAE C., KIM B. & HEO J. (2020). EdNet : A Large-Scale Hierarchical Dataset in Education. In I. I. BITTENCOURT, M. CUKUROVA, K. MULDNER, R. LUCKIN & E. MILLÁN, Éds., *Artificial Intelligence in Education*, volume 12164, p. 69–73. Cham : Springer International Publishing. DOI : 10.1007/978-3-030-52240-7_13.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

GEORGESCU A. A. (2018). CHATBOTS FOR EDUCATION - TRENDS, BENEFITS AND CHALLENGES. p. 195–200, Bucharest, RO. DOI : 10.12753/2066-026X-18-097.

HOGAN A., BLOMQVIST E., COCHEZ M., D'AMATO C., MELO G. D., GUTIERREZ C., KIRRANE S., GAYO J. E. L., NAVIGLI R., NEUMAIER S., NGOMO A.-C. N., POLLERES A., RASHID S. M., RULA A., SCHMELZEISEN L., SEQUEDA J., STAAB S. & ZIMMERMANN A. (2022). Knowledge Graphs. *ACM Computing Surveys*, **54**(4), 1–37. DOI : 10.1145/3447772.

JIANG F. (2024). Identifying and mitigating vulnerabilities in llm-integrated applications. Mémoire de master, University of Washington.

KUMAR P. (2024). Large language models (LLMs) : survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, **57**(10), 260. DOI : 10.1007/s10462-024-10888-y.

LE THUAUT M. (2005). Le boom des formations en alternance.

LUO L., LI Y.-F., HAFFARI G. & PAN S. (2023). Reasoning on graphs : Faithful and interpretable large language model reasoning. *arXiv preprint arXiv :2310.01061*.

MINISTÈRE DU TRAVAIL, DE LA SANTÉ D. S. E. D. F. (2024). Les chiffres de l'apprentissage en 2022.

PÉREZ J. Q., DARADOUMIS T. & PUIG J. M. M. (2020). Rediscovering the use of chatbots in education : A systematic literature review. *Computer Applications in Engineering Education*, **28**(6), 1549–1565. DOI : 10.1002/cae.22326.

QU K., LI K. C., WONG B. T. M., WU M. M. F. & LIU M. (2024). A Survey of Knowledge Graph Approaches and Applications in Education. *Electronics*, **13**(13), 2537. DOI : 10.3390/electronics13132537.

RAMANDANIS D. & XINOGALOS S. (2023). Investigating the Support Provided by Chatbots to Educational Institutions and Their Students : A Systematic Literature Review. *Multimodal Technologies and Interaction*, **7**(11), 103. DOI : 10.3390/mti7110103.

STANLEY T. & XU J. (2019). Work-Integrated Learning in accountancy at Australian universities – forms, future role and challenges. *Accounting Education*, **28**(1), 1–24. DOI : 10.1080/09639284.2018.1454333.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A., BABAEI Y., BASHLYKOV N., BATRA S., BHARGAVA P., BHOSALE S. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models. *arXiv preprint arXiv :2307.09288*.

UCLARAY A., MAGDASOC T. M., NOORA A. J. & SALES M. V. T. (2023). Social Work Students' Challenges in Flexible Learning and Implications for Social Work Education : A Study in Bicol, Philippines. *Asean Social Work Journal*, **11**(1), 13–27. DOI : 10.58671/aswj.v11i1.36.

UNIVERSITY OF NOTRE DAME, AUSTRALIA, DOOLAN M., PIGGOTT B., UNIVERSITY OF NOTRE DAME, AUSTRALIA, CHAPMAN S., UNIVERSITY OF NOTRE DAME, AUSTRALIA, RYCROFT P. & UNIVERSITY OF NOTRE DAME, AUSTRALIA (2019). The Benefits and Challenges of Embedding Work Integrated Learning : A Case Study in a University Education Degree Program. *Australian Journal of Teacher Education*, **44**(6), 91–108. DOI : 10.14221/ajte.2018v44n6.6.

WAN Z., WANG X., LIU C., ALAM S., ZHENG Y., LIU J., QU Z., YAN S., ZHU Y., ZHANG Q., CHOWDHURY M. & ZHANG M. (2024). Efficient Large Language Models : A Survey. arXiv :2312.03863, DOI : 10.48550/arXiv.2312.03863.

WANG B. (2021). Mesh-Transformer-JAX : Model-Parallel Implementation of Transformer Language Model with JAX. https://github.com/kingoflolz/mesh-transformer-jax.

WANG Y., WANG M., MANZOOR M. A., LIU F., GEORGIEV G. N., DAS R. J. & NAKOV P. (2024). Factuality of Large Language Models : A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 19519–19529, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.1088.

WOLLNY S., SCHNEIDER J., DI MITRI D., WEIDLICH J., RITTBERGER M. & DRACHSLER H. (2021). Are we there yet ?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, **4**, 654924.

WORKSHOP B., SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Une approche hybride de l'IA pour les technologies éducatives : augmenter les STI avec l'IA générative

Sofiya Kobylyanskaya[1] Catherine de Vulpillières[1] Pierre-Yves Oudeyer[1, 2]
(1) EvidenceB, Paris, France
(2) Inria, Bordeaux, France
sofiya-k@evidenceb.org catherine-d@evidenceb.com
pierre-yves.oudeyer@inria.fr

RÉSUMÉ
Nous proposons une approche hybride de l'IA au service de l'éducation, en combinant la personnalisation offerte par les Systèmes de Tutorat Intelligents (STI) avec de l'IA générative permettant de générer un grand nombre de contenus éducatifs de qualité, tout en respectant les contraintes pédagogiques et cognitives.

ABSTRACT
**A hybrid AI approach to educational technologies : augmenting ITS with generative AI**

We propose a hybrid approach to AI for education, combining the personalization offered by Intelligent Tutoring Systems (ITS) with generative AI to generate a large number of high-quality educational contents, while respecting pedagogical and cognitive constraints.

MOTS-CLÉS : STI, personnalisation, approche hybride, IA générative.

KEYWORDS: ITS, personnalization, hybrid approach, generative AI.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

## 1 Contexte du projet

Le déclin des performances scolaires, mis en évidence par les évaluations PISA (Programme for International Student Assessment) et TIMSS (Trends in Mathematics and Science Study), souligne l'urgence de solutions éducatives efficaces, notamment pour les élèves défavorisés. Les outils numériques, et en particulier les systèmes d'apprentissage adaptatif, offrent des perspectives prometteuses en personnalisant l'enseignement selon les profils des apprenants. Les Systèmes de Tutorat Intelligents (STI), fondés sur des recherches en sciences cognitives, ont prouvé leur efficacité pour l'apprentissage de concepts complexes comme les mathématiques. Toutefois, la plupart de ces systèmes n'exploitent pas encore pleinement les possibilités offertes par l'IA générative.

En même temps, l'essor des modèles de langage de grande taille (LLM) transforme les pratiques éducatives, avec une adoption massive par les jeunes. Bien que les LLMs montrent un potentiel pour le tutorat, leur usage reste limité par des lacunes pédagogiques, une personnalisation insuffisante et l'absence de jeux de données éducatifs adaptés (Jurenka *et al.*, 2024; Tack & Piech, 2022). En plus, l'usage massif de tuteurs basés entièrement sur l'IA générative pose des questions d'impact

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 145–148.

environnemental. Certains projets, comme LearnLM [1] (Google) ou MathDIAL (ETH Zurich) (Macina *et al.*, 2023), explorent l'usage de l'IA générative sur des données éducatives, mais soulignent le manque d'ensembles de données ouvertes sur les interactions élèves-enseignants. Quant aux STIs, bien que leur efficacité soit bien établie quand ils sont utilisés de manière adéquate (Clément *et al.*, 2015, 2025; Kulik & Fletcher, 2016; Létourneau *et al.*, 2025), elle repose sur une production manuelle coûteuse de contenus.

Nous proposons une approche hybride, combinant la génération automatique de ressources pédagogiques par les LLMs avec l'orchestration fine des STIs, et l'utilisation future de modèles de petite taille (SLM) pour une adaptation en temps réel aux besoins des élèves.

## 2 Méthodologie

EvidenceB propose des STIs visant l'apprentissage de la littératie et de la numératie, tels qu'Adaptiv'Math [2] et Adaptiv'Langue [3]. Notre méthodologie repose sur une combinaison de plusieurs formes d'intelligence artificielle ayant pour objectif de renforcer l'apprentissage. L'architecture des STIs d'EvidenceB s'appuie sur un graphe d'exercices structuré par des experts pédagogiques (Clément *et al.*, 2015, 2025). Ce graphe est le support d'algorithmes de personnalisation qui orchestrent des parcours adaptés aux besoins des élèves. Une fois ces parcours réalisés, les élèves sont regroupés à l'aide d'un algorithme de clustering, permettant aux enseignants de visualiser des profils d'apprentissage dans un tableau de bord.

Notre objectif est d'enrichir cette architecture en y intégrant l'IA générative pour produire automatiquement une diversité d'exercices respectant à la fois la structure du graphe et les contraintes pédagogiques et cognitives. À chaque étape — conception, validation, déploiement — l'expertise humaine demeure centrale pour assurer la qualité des contenus.

La première phase consiste à concevoir les exercices à l'aide de LLM et d'experts. Des spécialistes en sciences cognitives et pédagogues conçoivent des exercices-types et définissent les contraintes (objectifs, difficulté, compétences) qui servent de référence aux LLMs pour générer des exercices conformes aux exigences pédagogiques.

La deuxième phase porte sur la validation des contenus générés, selon des critères tels que clarté, validité pédagogique, difficulté, cohérence et vraisemblance. Cette validation est réalisée à la fois par des experts humains et par le LLM lui-même en tant qu'évaluateur automatisé (*LLM-as-judge*). La fiabilité du LLM est évaluée via l'alt-test (Calderon *et al.*, 2025), en comparant ses jugements à ceux des experts. Le LLM a pour rôle d'effectuer un premier filtrage des exercices générés sur la base des critères établis, tandis que les annotateurs humains valident ces exercices selon les mêmes critères, tout en ajustant les contraintes pédagogiques et cognitives à prendre en compte par le LLM afin d'améliorer la pertinence et la fiabilité des afin d'améliorer la pertinence et la fiabilité des futures générations et évaluationsfutures générations et évaluations.

Les exercices validés sont ensuite organisés sous forme de graphes et intégrés dans le STI. Ce système repose sur un algorithme d'apprentissage par renforcement (ZPDES, un bandit multi-bras (Clément *et al.*, 2015) ) — conçu pour personnaliser le parcours d'apprentissage des élèves. L'algorithme ajuste dynamiquement le parcours de chaque élève dans le graphe en fonction de ses performances

1. https://ai.google.dev/gemini-api/docs/learnlm?hl=fr
2. https://evidenceb.fr/produits/adaptiv-math/
3. https://evidenceb.fr/produits/adaptiv-langue/

dans le but de maximiser le progrès d'apprentissage (Gottlieb & Oudeyer, 2018). En proposant des exercices adaptés au niveau et à l'évolution de chaque élève, il construit ainsi un parcours personnalisé, continuellement mis à jour pour refléter les besoins et les acquis de l'élève.

Au fur et à mesure de la progression, un algorithme de clustering (K-means) regroupe les élèves selon plusieurs critères : les scores obtenus à chaque exercice, le nombre d'exercices réalisés par module, le nombre d'objectifs ouverts par module, ainsi que le temps de réponse. Ces regroupements, mis à jour dynamiquement à chaque avancée des élèves dans leur parcours, sont affichés dans le tableau de bord enseignant (mais restent invisibles pour les élèves). Ils permettent de faciliter le suivi individualisé, de repérer plus facilement les difficultés, et d'organiser, si besoin, des groupes de remédiation. Les groupes sont indicatifs et non définitifs. Ils ne sont pas labellisés, car ils ne reflètent pas directement un niveau (plus ou moins fort) mais regroupent des élèves dont les parcours présentent des similarités.

Cette approche hybride exploite les complémentarités entre plusieurs formes d'IA : elle combine la robustesse des STIs pour personnaliser les parcours d'apprentissage, avec la capacité de l'IA générative à produire un grand volume d'exercices, tout en maintenant l'intervention humaine afin de garantir l'efficacité et la fiabilité. Par ailleurs, cette méthode se révèle plus économe en consommation énergétique par rapport aux méthodes reposant exclusivement sur des LLMs.

La pertinence de cette approche sera évaluée en deux temps : (1) une validation de la qualité des exercices générés, par des experts et via le *LLM-as-judge* ; (2) une étude d'impact sur le terrain, sous forme d'un essai randomisé contrôlé (RCT) (Roell *et al.*, 2025a,b), mesurant les effets sur les compétences (pré/post-tests), ainsi que sur la motivation et l'engagement via l'analyse des retours utilisateurs.

# 3 Limites et discussion

Nous proposons une approche hybride combinant différents types d'intelligence artificielle, dont les LLMs, afin de générer automatiquement un grand nombre d'exercices répondant à des spécifications pédagogiques précises. Ces exercices sont ensuite utilisés pour la personnalisation de l'apprentissage.

Cette démarche nécessite toutefois une prise en compte rigoureuse des limites actuelles des LLMs. N'étant pas spécifiquement entraînés sur des corpus éducatifs, ces modèles peuvent produire des exercices qui ne correspondent pas aux objectifs pédagogiques visés, sont inadaptés au niveau attendu, ou encore comportent des erreurs factuelles (par exemple, une mauvaise application de formule), des formulations ambiguës, ou des contextes peu pertinents sur le plan didactique. En particulier, les LLMs peinent notamment à structurer des questions selon les exigences pédagogiques, et à proposer une correction à la fois juste, explicite et didactique.

Pour garantir la qualité des contenus générés, l'intervention humaine reste essentielle à plusieurs niveaux. L'usage des LLMs est strictement encadré : ils interviennent uniquement pour la génération de contenu, selon des contraintes formulées par des experts. Ils n'interviennent ni dans la conception pédagogique (choix des objectifs, des sujets, etc.), ni dans la personnalisation des parcours des élèves. Enfin, tous les exercices produits sont systématiquement validés par des experts avant leur intégration dans le STI.

# Références

CALDERON N., REICHART R. & DROR R. (2025). The alternative annotator test for LLM-as-a-

judge : How to statistically justify replacing human annotators with llms. arXiv : 2501.10970.

CLÉMENT B., ROY D., OUDEYER P.-Y. & LOPES M. (2015). Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining*, **7**(2). arXiv : 1310.3174, DOI : 10.5281/zenodo.3554667.

CLÉMENT B., SAUZÉON H., ROY D. & OUDEYER P.-Y. (2025). Improved performances and motivation in intelligent tutoring systems : Combining machine learning and learner choice. arXiv : 2402.01669.

GOTTLIEB J. & OUDEYER P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, **19**(12), 758–770. DOI : 10.1038/s41583-018-0078-0.

JURENKA I., KUNESCH M., MCKEE K. R., GILLICK D., ZHU S., WILTBERGER S., PHAL S. M., HERMANN K., KASENBERG D., BHOOPCHAND A., ANAND A., PÎSLAR M., CHAN S., WANG L., SHE J., MAHMOUDIEH P., RYSBEK A., KO W.-J., HUBER A., WILTSHIRE B., ELIDAN G., RABIN R., RUBINOVITZ J., PITARU A., MCALLISTER M., WILKOWSKI J., CHOI D., ENGELBERG R., HACKMON L., LEVIN A., GRIFFIN R., SEARS M., BAR F., MESAR M., JABBOUR M., CHAUDHRY A., COHAN J., THIAGARAJAN S., LEVINE N., BROWN B., GORUR D., GRANT S., HASHIMSHONI R., WEIDINGER L., HU J., CHEN D., DOLECKI K., AKBULUT C., BILESCHI M., CULP L., DONG W.-X., MARCHAL N., DEMAN K. V., MISRA H. B., DUAH M., AMBAR M., CACIULARU A., LEFDAL S., SUMMERFIELD C., AN J., KAMIENNY P.-A., MOHDI A., STRINOPOULOUS T., HALE A., ANDERSON W., COBO L. C., EFRON N., ANANDA M., MOHAMED S., HEYMANS M., GHAHRAMANI Z., MATIAS Y., GOMES B. & IBRAHIM L. (2024). Towards responsible development of generative ai for education : An evaluation-driven approach. arXiv : 2407.12687.

KULIK J. A. & FLETCHER J. D. (2016). Effectiveness of intelligent tutoring systems : a meta-analytic review. *Review of educational research*, **86**(1), 42–78. DOI : 10.3102/0034654315581420.

LÉTOURNEAU A., DESLANDES MARTINEAU M., CHARLAND P., KARRAN J. A., BOASEN J. & LÉGER P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. *npj Science of Learning*, **10**(1), 1–13. DOI : 10.1038/s41539-025-00320-7.

MACINA J., DAHEIM N., CHOWDHURY S., SINHA T., KAPUR M., GUREVYCH I. & SACHAN M. (2023). MathDial : A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 5602–5621, Singapore : Association for Computational Linguistics.

ROELL M., DE VULPILLIÈRES C., KNOPS A. & VAGHARCHAKIAN L. (2025a). From intuition to abstraction : Supporting the transition to formal fraction understanding with AI-powered tools. In *The Mathematical Cognition and Learning Society (MCLS)*, Hong Kong. In I. Resnick (Chair), The malleability and utility of informal fraction knowledge from early years contexts through formal schooling.

ROELL M., DE VULPILLIÈRES C., KNOPS A. & VAGHARCHAKIAN L. (2025b). Leveraging adaptive digital tools to enhance early mathematics learning : Insights from randomized controlled trials. In *Conference for Research in Early Childhood Education (CRECE 2025)*, Hong Kong. In V. Simms (Chair), Unlocking Math Potential : Breakthrough Strategies for All Ages.

TACK A. & PIECH C. (2022). The AI teacher test : Measuring the pedagogical ability of Blender and GPT-3 in educational dialogues. In A. MITROVIC & N. BOSCH, Éds., *Proceedings of the 15th International Conference on Educational Data Mining*, p. 522–529, Durham, United Kingdom : International Educational Data Mining Society. DOI : 10.5281/zenodo.6853187.

*Actes de CORIA-TALN-RJCRI-RECITAL 2025*
*Marseille, France, 30 juin au 4 juillet 2025*

# Vers des RAGs intégrant véracité, subjectivité et explicabilité

Alae Bouchiba[1] Adrian-Gabriel Chifu[1] Sébastien Fournier[1] Lorraine Goeuriot[2]
Philippe Mulhem[2]
(1) Aix Marseille Univ, CNRS, LIS, Marseille, France
(2) Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP[*], LIG, 38000 Grenoble, France
(1) {prénom}.{nom}@lis-lab.fr, (2) {prénom}.{nom}@univ-grenoble-alpes.fr

RÉSUMÉ

Cet article introduit X-RAG-VS , un cadre pour intégrer véracité , subjectivité et explicabilité dans les systèmes RAG , en réponse aux besoins éducatifs. À travers des cas d'usage et l'analyse de modèles existants , nous montrons que ces dimensions restent insuffisamment prises en compte. Nous proposons une approche unifiée pour des réponses plus fiables , nuancées et explicables.

ABSTRACT

**Towards RAGs Integrating Veracity, Subjectivity, and Explainability**

We introduce X-RAG-VS , a framework aiming to integrate veracity , subjectivity , and explainability within retrieval-augmented generation systems for educational use. Based on concrete use cases and an analysis of current large language models , we show that these critical dimensions are often addressed separately or insufficiently. We propose a unified approach to support more reliable , nuanced, and transparent AI-generated content.

MOTS-CLÉS : RAG , éducation , véracité, subjectivité , pensée critique , explicabilité.

KEYWORDS: RAG , education , veracity , subjectivity , critical thinking , explainability.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

## 1 Introduction

Cet article propose l'étude des systèmes de génération augmentée par récupération (RAG) (Lewis *et al.*, 2020) adaptés aux besoins éducatifs , ayant pour objectif de favoriser l'esprit critique , la transparence et la pluralité des points de vue. Nous articulons nos travaux au travers des trois dimensions clés suivantes : **véracité**, **subjectivité** et **explicabilité**. Nous nous concentrons sur ces trois dimensions, nécessaires pour garantir des réponses à la fois fiables , nuancées et accessibles.

Nous proposons ainsi les fondements d'un cadre méthodologique, que nous appelons X-RAG-VS , visant à explorer comment ces dimensions , souvent traitées séparément dans la littérature , peuvent être combinées de manière cohérente pour répondre à des enjeux pédagogiques concrets : aider les apprenants à juger de la fiabilité d'une information, à confronter des opinions divergentes et à saisir les mécanismes de raisonnement d'un système d'intelligence artificielle (IA).

∗. Institute of Engineering Univ. Grenoble Alpes

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 149–155.

Cette contribution s'inscrit en tant que réflexion structurée sur les conditions d'un usage responsable de l'IA générative en contexte éducatif. Elle se positionne comme un travail exploratoire , à la croisée entre cadre théorique et mise en perspective empirique , destiné à nourrir les discussions sur la formation à l'esprit critique à l'ère des grands modèles de langage.

L'analyse qui suit est décomposée en 3 étapes : tout d'abord nous allons décrire un exemple lié au domaine de l'éducation, pour lequel les 3 dimensions sont nécessaires conjointement. Tout comme dans (Chen *et al.*, 2023), on peut se poser la question de savoir dans quelle direction faire porter le travail de recherche : étudier les prompts ou bien modifier les modèles LLMs. Dans un second temps, nous allons montrer que ces dimensions sont grandement absentes dans les réponses générées par deux modèles de chat connus , au travers de prompts simples. Nous allons ensuite explorer dans quelle mesure les approches de la littérature sont capables d'intégrer tous ces éléments. Nous proposerons ensuite un plan de travail pour proposer de dépasser les limites existantes.

## 2 Cas d'usage et axes d'analyse

Afin d'analyser les besoins et l'existant pour nos trois dimensions d'étude , nous définissons dans la suite un cadre d'analyse basé sur des cas d'usage.

Les cas d'usage sont , pour nous , définis : i) par des étudiants ou des écoliers , ii) dans le cadre de la recherche d'un panorama d'avis sur un sujet de société. Ces cas d'usage se situent plus globalement dans l'"éducation à la pensée critique" : cette dernière repose sur des arguments et des faits et sur leur mise en perspective argumentée, qui eux-mêmes reposent donc sur des éléments pour lesquels la véracité est caractérisée (est-ce que les arguments sont vrais) , pour lesquels la subjectivité/objectivité est estimée (est-ce que les éléments sont objectifs ou subjectifs) , et enfin pour lesquels des explications sur les raisons de la présentation des éléments sont fournies (pourquoi ces arguments sont présentés).

Plus précisément :

— **la véracité** est entendue comme la capacité à produire des informations exactes (Hwang *et al.*, 2024) et fondées sur des sources fiables (Zhou *et al.*, 2024b) ;

— **la subjectivité** , qui renvoie à la prise en compte de points de vue divergents (Wan & McAuley, 2016), d'opinions ou de jugements de valeur (Balayn & Bozzon, 2019) ;

— **l'explicabilité** , c'est-à-dire la faculté pour le système d'expliquer de manière intelligible les choix réalisés, les sources mobilisées ou les raisonnements suivis (Arrieta *et al.*, 2020).

Ces dimensions sont aujourd'hui considérées comme essentielles pour concevoir des systèmes d'IA plus transparents, nuancés et pédagogiquement adaptés (Maity & Deroy, 2024) et (Karran *et al.*, 2024).

Le succès des LLMs pour nos cas d'usage pourraient grandement bénéficier de la prise en compte **conjointement** de ces trois dimensions : les textes générés seraient alors par exemple capables d'expliquer pourquoi (explicabilité) un point de vue qui est supporté par des informations vraies (véracité) mais supportées par des sources d'information subjectives (subjectivité) est choisi pour être présenté dans l'argumentaire de la réponse, car ce point de vue est l'un des éléments qui permet d'expliquer la complexité d'un sujet.

# 3 Utilisation de prompts pour les trois dimensions explorées

Cette section a pour objectif d'estimer dans quelle mesure certains LLMs de chat intègrent par défaut certains éléments des dimensions de véracité, de subjectivité et d'explicabilité.

L'idée d'explorer comment les prompts seraient capables d'intégrer la prise en compte de ces trois dimensions vient immédiatement à l'esprit, car il a été montré (Zhou *et al.*, 2024a) que le prompt-tuning a un impact important sur la qualité des réponses de ces modèles sans nécessiter d'apprentissage. Afin d'évaluer empiriquement cette hypothèse , cette étude propose une expérimentation originale utilisant deux LLMs dédiés aux *chat* , ChatGPT 4o en mode "recherche web" et Gemini 2.5 Pro. Bien qu'ils ne soient pas des RAG , ils sont capables de sourcer (i.e., indiquer des références) leur réponse , ce qui nous permet d'étudier leur comportement par rapport à nos dimensions d'intérêt. Notre étude vise à analyser , pour un contexte d'utilisation spécifique en se basant sur des prompts simples , si et comment ces systèmes sont capables d'intégrer ces trois dimensions de manière satisfaisante.

Les deux requêtes soumises dans cette étude correspondent au niveau 1 de la taxonomie TELeR (Santu & Feng, 2023) , le choix de celles-ci s'est basé sur le choix de l'utilisateur , pour délibérément imiter ce qu'un collégien ou un lycéen les formuleraient. Nous testons la capacité du système à explorer les dimensions naturellement sans pour autant complexifier la requête . Le choix des deux requêtes était : « Faut-il interdire les devoirs à la maison ?» et «Faut-il raccourcir les vacances d'été ?».

Dans cette perspective, nous avons analysé les réponses générées par ces deux grands modèles de langues.

Les deux modèles de langues utilisés présentent une analogie structurée dans leur manière d'aborder le sujet ; ils adoptent la même posture argumentative , exposant de façon équilibrée les points favorables et défavorables à la question posée.

L'analyse détaillée exhibe cependant les éléments suivants :

— La symétrie des arguments dans la présentation des idées donne l'illusion d'une neutralité, mais masque en réalité une absence d'engagement critique et ne laisse pas la place à une véritable pluralité de points de vue vécus , ni à l'expression des différentes positions incarnées.

— Le choix des références mobilisées, souvent issues de blogs ou de sites web non académiques, renforce cette impression de surface argumentative : les contenus sont illustratifs , mais ne s'appuient pas sur des sources clairement identifiées comme fiables.

— Le raisonnement sous-jacent et le processus de sélection des sources sont peu ou pas explicités dans les réponses générées. Cette opacité rend difficile l'identification des parties du texte sur lesquelles le modèle s'appuie réellement pour construire son argumentation , et empêche de comprendre comment il réagit, structure sa réflexion, ou mobilise des connaissances face à la question posée. On a donc aucune information d'explication fournie par ces modèles.

Aucun des deux modèles testés ne parvient à intégrer de manière cohérente et satisfaisante l'ensemble des dimensions ciblées. Si certains éléments , comme la structuration explicative ou la présence d'arguments opposés , sont prometteurs , ils restent insuffisants en l'absence de vérification documentaire rigoureuse et de véritable diversité argumentative.

On pourrait arguer que des prompts plus complexes seraient peut-être à même d'améliorer la prise en compte de nos dimensions d'analyse. IL a cependant été montré récemment (Zou *et al.*, 2023).

la forte sensibilité des LLMs à la formulation des prompts , ainsi que les risques d'hallucinations persistantes dans des contextes sensibles (Dahl *et al.*, 2024). Ces observations confirment selon nous que le prompt-tuning [1] seul ne constitue pas une solution robuste pour les enjeux complexes que nous visons.

Ces résultats confirment alors , dans notre cas d'usage , que les trois dimensions identifiées comme essentielles à la pensée critique sont très inégalement prises en compte dans les modèles actuels . Cela rejoint les constats établis dans la littérature scientifique , que nous présentons dans la section suivante.

# 4 Les dimensions d'analyse dans la littérature

Dans cette section , nous examinons dans quelle mesure la littérature scientifique actuelle propose des réponses à notre problématique , en explorant des travaux portant explicitement sur la véracité , la subjectivité ou l'explicabilité dans des systèmes similaires.

## 4.1 Véracité

La véracité ne se réduit pas à la cohérence interne des réponses générées : elle implique la capacité du système à produire une information exacte , fondée sur des sources identifiables. Cette exigence est d'autant plus cruciale dans un cadre éducatif , où le développement de la pensée critique suppose que l'apprenant puisse évaluer la qualité des connaissances mobilisées. tout comme l'étude VeraCT (Chen *et al.*, 2024) qui affirme cette nécessité et propose une approche de vérification des faits par récupération et génération , dans laquelle chaque affirmation est confrontée à des sources externes crédibles , accompagnées d'un raisonnement structuré. Dans une perspective plus large , la synthèse (Zhao *et al.*, 2024) met en évidence les différentes stratégies de vérification de revendications intégrées aux LLMs , soulignant que la véracité constitue un pilier fondamental pour juger de la robustesse d'un système. Ce principe reste le même pour d'autres travaux, tout en évaluant la fiabilité des sources web à partir de la qualité factuelle des informations qu'elles véhiculent (Dong *et al.*, 2015) . Ces approches convergent vers une même conclusion : pour qu'un système RAG contribue réellement à l'apprentissage , il doit offrir un accès structuré à une information vérifiée , bien que traçable.

## 4.2 Subjectivité

Dans un contexte éducatif et de formation à l'esprit critique , il est essentiel que les systèmes soient capables de restituer non seulement des faits vérifiés , mais aussi la diversité des points de vue présents sur un sujet. La capacité à exposer des opinions divergentes est particulièrement cruciale dans les situations de débat ou d'analyse de controverses , où il n'existe pas une vérité unique mais plusieurs interprétations légitimes. Pourtant, cette dimension de subjectivité reste largement absente des architectures RAG actuelles. Des travaux récents , comme pour (Chen *et al.*, 2024) , mettent en évidence la nécessité d'intégrer des perspectives multiples pour enrichir la génération de contenu. De même , l'approche Vendi-RAG proposée (Rezaei & Dieng, 2025) introduit un mécanisme d'optimisation adaptative entre diversité et qualité , permettant de générer des réponses plus nuancées et représentatives de la pluralité des opinions. Dans le cadre éducatif , intégrer la subjectivité aux côtés de la véracité et de l'explicabilité est indispensable pour développer des outils qui ne se contentent

1. Le fait de modifier les prompts pour affiner la réponse, de manière experte.

pas de transmettre une information correcte , mais qui permettent aux apprenants de comprendre la complexité des débats , de comparer les arguments , et de forger leur propre jugement critique.

### 4.3 Explicabilité

l'explicabilité ne se limite pas à exposer les documents utilisés, mais implique la capacité du système RAG à articuler de manière intelligible le raisonnement qui relie ces sources à la réponse générée , cette forme d'explicabilité est particulièrement précieuse dans le cadre éducatif , où la compréhension du raisonnement importe autant que le résultat produit. Dans ce sens , (Xu *et al.*, 2023) proposent une approche de RAG interprétable , dans laquelle chaque segment de réponse est justifié par un élément explicite du document source , renforçant ainsi la traçabilité et la vérifiabilité de l'information générée. De manière complémentaire , (Tian *et al.*, 2023) introduisent un pipeline de génération structuré où les sources récupérées sont non seulement affichées , mais réorganisées dans un graphe d'argumentation destiné à expliciter la logique suivie par le modèle. Ces efforts montrent que l'explicabilité dans RAG ne peut être atteinte uniquement par la transparence des sources , mais nécessite également une mise en forme narrative et rationnelle de l'information.

D'autres travaux visent également à combiner plusieurs de nos dimensions clés , comme pour (Abolghasemi *et al.*, 2024) qui s'intéressent à la manière dont les biais d'attribution influencent à la fois la fiabilité des sources mobilisées et la pluralité des perspectives exposées, articulant ainsi véracité et subjectivité. Par ailleurs , (Rezaei & Dieng, 2025) propose un cadre adaptatif qui équilibre diversité argumentative et cohérence des réponses générées , contribuant à une meilleure structuration du contenu et à une forme d'explicabilité implicite , sans pour autant rassembler les trois.

- Notre analyse ci-dessus révèle ainsi des propositions qui sont fragmentées , car chacune des dimensions qui nous intéressent sont traitées de manière indépendante , ou dépendante mais incomplète. Nous défendons le point de vue qu'une approche unifiée serait plus à même d'être capable d'articuler conjointement les dimensions de véracité , d'explicabilité et de subjectivité dans une même architecture.

- Il convient de noter que bien que les trois dimensions de véracité , subjectivité et explicabilité puissent présenter des interdépendances. Par exemple , l'explicabilité peut contribuer à éclairer la fiabilité des sources (véracité) ou la diversité des perspectives présentées (subjectivité). Cette analyse les considère comme des dimensions distinctes et autonomes. Cette approche méthodologique permet d'identifier plus précisément les lacunes spécifiques à chaque dimension dans les systèmes actuels et de concevoir des solutions ciblées pour chacune d'entre elles , avant d'envisager leur intégration cohérente dans un cadre unifié. De plus , cette distinction analytique facilite l'évaluation comparative des différentes approches de la littérature et permet de mieux cerner les contributions spécifiques de chaque travail par rapport aux trois dimensions considérées. C'est précisément cette analyse différenciée qui révèle la nécessité d'un cadre comme X-RAG-VS , capable d'orchestrer ces dimensions de manière synergique plutôt que de les traiter comme des éléments isolés ou faiblement articulés.

## 5 Vers une intégration de la véracité, la subjectivité et l'explicabilité

Il apparaît , sans pour autant être des preuves définitives , d'après les quelques expérimentations menées , que ni la simple ingénierie de prompts , ni les approches existantes dans la littérature ne

permettent , à ce jour , de prendre en compte de manière intégrée les trois dimensions qui nous intéressent.

L'approche X-RAG-VS que nous proposons repose sur deux axes complémentaires. D'une part , elle intègre les dimensions de véracité , subjectivité et explicabilité directement dans le processus de génération , en agissant à un niveau sémantique fin. D'autre part , elle s'inspire de mécanismes tels que SELF-RAG (Liu, 2023) , qui déclenchent dynamiquement la recherche documentaire , adaptés ici à des finalités éducatives.

Une telle prend tout son sens dans des contextes d'apprentissage concrets. Par exemple , lorsqu'un élève cherche à se forger une opinion sur une question de société , la génération ne se limite pas à structurer une réponse : elle articule des informations vérifiables , présente une diversité de points de vue et explicite le raisonnement suivi. Le système devient alors un véritable support au développement de la pensée critique , en phase avec les cas d'usage identifiés.

## Remerciements

Cette recherche a été financée en partie par l'Agence nationale de la recherche (ANR) au titre du projet GUIDANCE , ANR-23-IAS1-0003.

## Références

ABOLGHASEMI A., AZZOPARDI L., HASHEMI S. H. *et al.* (2024). Evaluation of attribution bias in retrieval-augmented large language models. *arXiv preprint arXiv :2410.12380*.

ARRIETA A. B., DÍAZ-RODRÍGUEZ N., SER J. D., BENNETOT A., TABIK S., BARBADO A., GARCIA S., GIL-LOPEZ S., MOLINA D., BENJAMINS R., CHATILA R. & HERRERA F. (2020). Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, **58**, 82–115. arXiv :1910.10045.

BALAYN A. & BOZZON A. (2019). Designing evaluations of machine learning models for subjective inference : The case of sentence toxicity. arXiv preprint arXiv :1911.02471.

CHEN B., YI F. & VARRÓ D. (2023). Prompting or fine-tuning ? a comparative study of large language models for taxonomy construction. arXiv :2309.01715.

CHEN T., SORENSEN J., ZIEMS C. *et al.* (2024). Retrieval-augmented generation with diverse perspectives. *arXiv preprint arXiv :2409.18110*.

DAHL M., MAGESH V., SUZGUN M. & HO D. E. (2024). Large legal fictions : Profiling legal hallucinations in large language models. arXiv preprint arXiv :2401.01301.

DONG X. L., GABRILOVICH E., MURPHY K., DANG V., HORN W., LUGARESI C., SUN S. & ZHANG W. (2015). Knowledge-based trust : estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, **8**(9), 938–949. DOI : 10.14778/2777598.2777603.

HWANG S., BAEK J., PARK J. & KANG J. (2024). Retrieval-augmented generation with estimation of source reliability. https://arxiv.org/abs/2410.22954. arXiv :2410.22954.

KARRAN T., HALL M. & NUNAN T. (2024). Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education. *arXiv preprint arXiv :2402.15027*.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., TAU YIH W., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. NeurIPS 2020.

LIU Z. E. A. (2023). Self-rag : Retrieval-augmented generation via self-retrieval. https://arxiv.org/abs/2310.11511. arXiv preprint arXiv :2310.11511v1.

MAITY D. & DEROY Y. (2024). Human-centric explainable ai in education. *arXiv preprint arXiv :2410.19822*.

REZAEI M. R. & DIENG A. B. (2025). Vendi-rag : Adaptively trading-off diversity and quality significantly improves retrieval augmented generation with llms. *arXiv preprint arXiv :2502.11228*.

SANTU S. K. K. & FENG D. (2023). Teler : A general taxonomy of llm prompts for benchmarking complex tasks. *arXiv preprint arXiv :2305.11430*.

TIAN Y., YE D., LIN Y., LIU Z. & SUN M. (2023). Explanation graph generation via pre-trained language models. *arXiv preprint arXiv :2305.14277*.

WAN M. & MCAULEY J. (2016). Modeling ambiguity, subjectivity, and diverging viewpoints in opinion question answering systems. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, p. 489–498 : IEEE. DOI : 10.1109/ICDM.2016.0060.

XU C., YU M., GAO J., GAO Z., LIN J. & CALLAN J. (2023). Towards interpretable retrieval-augmented generation : A case study on explainable qa. *arXiv preprint arXiv :2310.03667*.

ZHAO L., LIU X. & LIU Q. (2024). Claim verification in the age of large language models : A survey. *arXiv preprint arXiv :2408.14317*.

ZHOU X., BEHROOZ M., DEHGHANI M. & REN X. (2024a). The prompt report : A systematic survey of prompting techniques. arXiv preprint arXiv :2406.06608v2.

ZHOU Y., LIU Y., LI X., JIN J., QIAN H., LIU Z., LI C., DOU Z., HO T.-Y. & YU P. S. (2024b). Trustworthiness in retrieval-augmented generation systems : A survey. arXiv :2409.10102.

ZOU A., WANG Z., CARLINI N., NASR M., KOLTER J. Z. & FREDRIKSON M. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv :2307.15043*.