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

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.