GAIMHE : hybridation des systèmes tutoriels intelligents adaptatifs et de l'IA générative pour l'éducation

Sofiya Kobylyanskaya[1] Olivier Clerc[2] Axelle Ziegler[1] Catherine de Vulpillières[1] Pierre-Yves Oudeyer[1,2]

(1) EvidenceB, Paris, France

(2) Centre Inria de l'Université de Bordeaux, Talence, France

sofiya-k@evidenceb.fr, olivier.clerc@inria.fr, axelle-z@evidenceb.fr, catherine-d@evidenceb.fr, pierre-yves.oudeyer@inria.fr

RÉSUMÉ

Nous présentons GAIMHE (Generative AI and Hybrid Models for Education), un projet industriel-académique consacré à l'hybridation des systèmes tutoriels intelligents adaptatifs et de l'IA générative pour l'éducation. Le projet part d'un double constat : les systèmes adaptatifs permettent de personnaliser les parcours, mais leur extension reste coûteuse du fait de la production manuelle de contenus ; à l'inverse, les grands modèles de langage permettent une génération à grande échelle, tout en soulevant des enjeux de fiabilité, de contrôle pédagogique, de coût computationnel et d'intégrité des apprentissages. GAIMHE propose une architecture hybride dans laquelle le séquençage pédagogique reste gouverné par une architecture adaptative déjà déployée, tandis que l'IA générative est mobilisée, sous contraintes explicites, pour la pré-génération d'exercices, d'indices et de feedbacks. Ce résumé présente (i) le positionnement scientifique du projet, (ii) l'architecture envisagée, (iii) l'état d'avancement des pipelines de génération et de validation, et (iv) les principaux défis méthodologiques et pédagogiques liés à l'évaluation de cette hybridation en contexte scolaire réel.

ABSTRACT

GAIMHE : Hybridizing adaptive intelligent tutoring systems with generative AI for education

We present GAIMHE (Generative AI and Hybrid Models for Education), an industrial-academic project dedicated to the hybridization of adaptive intelligent tutoring systems and generative AI for education. The project starts from a twofold observation : adaptive systems support personalized learning paths, yet remain costly to scale because of manual content authoring ; conversely, large language models enable large-scale generation, but raise issues of reliability, pedagogical control, computational cost, and learning integrity. GAIMHE proposes a hybrid architecture in which pedagogical sequencing remains governed by an already deployed adaptive ITS architecture, while generative AI is used under explicit constraints for the pre-generation of exercises, hints, and feedback. This summary presents (i) the project's scientific positioning, (ii) the proposed architecture, (iii) the current progress of the generation and validation pipelines, and (iv) the main methodological and pedagogical challenges involved in evaluating such hybridization in real classroom settings.

MOTS-CLÉS : IA en éducation, systèmes tutoriels intelligents, IA générative, personnalisation, évaluation.

KEYWORDS: AI in education, intelligent tutoring systems, generative AI, personalization, evaluation.

1 Positionnement du projet

L'intégration de l'IA générative dans l'éducation suscite un fort intérêt, mais soulève aussi des interrogations sur ses effets réels sur les apprentissages. Les grands modèles de langage facilitent la génération de contenus, l'interaction en langage naturel et la production de feedbacks personnalisés, mais leur usage non contraint peut entraîner des erreurs factuelles, une aide trop directive, une réduction de l'effort cognitif de l'élève ou un affaiblissement du contrôle de l'enseignant (Kasneci *et al.*, 2023; Bastani *et al.*, 2025; Jurenka *et al.*, 2024). À l'inverse, les systèmes tutoriels intelligents (STI), qui modélisent les compétences des élèves afin d'adapter les activités proposées, permettent un séquençage adaptatif et une évaluation rigoureuse (Anderson *et al.*, 1995; Abdelrahman *et al.*, 2023), mais leur extension reste limitée par le coût de production de corpus d'exercices, d'indices et de feedbacks de qualité.

C'est dans ce contexte que s'inscrit GAIMHE (Generative AI and Hybrid Models for Education), un projet qui articule les atouts des STI et de l'IA générative. Il s'appuie sur une architecture de STI adaptatif développée par EvidenceB et déjà déployée à grande échelle, notamment à travers les plateformes Adaptiv'Math et MIA Seconde, utilisées dans plus de 5 000 écoles en France et à l'étranger. Dans cet environnement, les contenus sont organisés en graphes pédagogiques structurés et le choix des exercices est piloté par un moteur de recommandation adaptatif de type bandit multi-bras visant à maximiser le progrès d'apprentissage (Clément *et al.*, 2015, 2024).

GAIMHE part de deux limites de cette approche : l'extension du contenu pédagogique repose encore sur une production experte manuelle, lente et coûteuse ; et, si le séquençage des activités est personnalisé, l'aide fournie pendant la résolution reste souvent générique et peu adaptée aux erreurs produites, alors que la littérature souligne l'importance d'indices progressifs et contingents aux difficultés de l'élève (Colliot *et al.*, 2024; Tricomi & DePasque, 2016). Le projet vise donc à pré-générer des ressources pédagogiques sous contraintes didactiques explicites, à concevoir une chaîne de validation hybride combinant pré-filtrage automatique et validation experte humaine, puis à évaluer les effets éducatifs de cette hybridation en contexte scolaire réel.

L'originalité de GAIMHE ne réside pas dans l'usage isolé de modèles génératifs, mais dans leur intégration fonctionnellement circonscrite au sein d'un environnement adaptatif déjà structuré, avec une attention portée à la validité pédagogique, à la frugalité et à l'évaluation en usage réel. Cette contribution se situe ainsi au stade d'un cadrage méthodologique et d'un état d'avancement du projet : elle présente l'architecture technique et pédagogique de GAIMHE, les pipelines de génération-validation en cours de consolidation, ainsi que le protocole prévu pour évaluer cette hybridation en conditions scolaires réelles.

2 Architecture proposée et état d'avancement

GAIMHE propose de préserver le noyau adaptatif déjà utilisé tout en y intégrant des modules génératifs selon une répartition entre niveau macro-pédagogique et niveau micro-pédagogique. La séparation entre génération hors ligne et intervention générative en temps réel répond également à un enjeu de frugalité. Un système adaptatif déployé à grande échelle doit traiter des volumes très importants d'interactions élèves-exercices ; un recours systématique à des grands modèles de langage pour générer chaque exercice, indice ou feedback en temps réel serait donc difficilement soutenable en termes de coût d'inférence, de latence et de consommation énergétique. GAIMHE privilégie pour

cette raison la pré-génération hors ligne, le pré-filtrage automatique avant validation humaine, et la limitation des appels génératifs en temps réel aux situations pédagogiquement justifiées. Cette frugalité pourra être évaluée par des indicateurs tels que le nombre d'appels aux modèles, le coût moyen par ressource produite, le temps de génération et de validation, la latence des éventuelles interventions en temps réel et la proportion de contenus éliminés ou révisés à chaque étape du pipeline.

Au niveau macro-pédagogique, les modèles génératifs sont utilisés hors ligne pour produire à grande échelle des exercices, des distracteurs, des indices et des feedbacks à partir de spécifications pédagogiques explicites et d'exercices-types produits par les experts. Chaque exercice produit constitue une variation de cet exercice-type, garantissant cohérence pédagogique et diversification suffisante pour limiter les effets de familiarisation. Les spécifications transmises au modèle lors de la génération sont fondées sur les intentions pédagogiques et incluent, par exemple, un objectif du graphe pédagogique, un type d'exercice attendu, une compétence ciblée, une difficulté approximative, ainsi que des contraintes sur les distracteurs, comme la présence d'erreurs plausibles correspondant à des conceptions erronées fréquentes.

Au niveau micro-pédagogique, la génération en temps réel est envisagée comme un mécanisme d'aide ciblée, déclenchée soit à l'initiative de l'élève lorsque celui-ci sollicite explicitement une aide, soit automatiquement par le système sur la base de signaux observables calculés à partir de l'interaction, tels que des réponses incorrectes successives, un temps de résolution dépassant un seuil défini ou une production ouverte difficile à catégoriser. La sortie générée est limitée à un feedback aligné avec l'objectif pédagogique, l'historique immédiat de l'élève et les principes d'étayage progressif, sans divulguer prématurément la réponse. Ces déclencheurs devront être calibrés empiriquement afin de limiter les interventions inutiles, de préserver l'effort cognitif de l'élève et d'éviter une dépendance excessive à l'aide générée.

Le projet dispose déjà de plusieurs briques opérationnelles, dont une chaîne de génération automatique d'exercices et une chaîne de pré-filtrage fondée sur un dispositif de *LLM-as-Judge*. Ce module n'a pas vocation à remplacer l'expertise humaine, mais à prioriser les contenus à relire, identifier les erreurs manifestes et réduire le coût de validation avant arbitrage expert. Le pré-filtrage repose sur un codebook explicite : pour les exercices, les critères portent sur la conformité au modèle fourni, la correction de l'énoncé et de la réponse, la clarté, l'adéquation au niveau scolaire, l'alignement avec l'objectif pédagogique, la plausibilité du contexte et la qualité des distracteurs ; pour les indices et feedbacks, ils portent sur la correction, la clarté, la suffisance de l'aide, l'absence de divulgation prématurée de la réponse et l'alignement avec les principes d'étayage progressif.

La calibration du *LLM-as-Judge* suit une procédure en plusieurs étapes. Dans un premier temps, un sous-ensemble de contenus générés est annoté par au moins trois annotateurs humains, qui sont des enseignants des disciplines concernées, à partir d'un codebook commun. Cette étape vise à vérifier la fiabilité de l'annotation humaine : l'accord inter-annotateurs est estimé par critère à l'aide du Krippendorff's Alpha (Hayes & Krippendorff, 2007) ; les désaccords identifiés servent à clarifier les consignes ou à réviser le codebook. Dans un second temps, le *LLM-as-Judge* est appliqué aux mêmes contenus que ceux annotés par les enseignants. La possibilité d'utiliser le *LLM-as-Judge* comme annotateur alternatif sera évaluée statistiquement à l'aide de l'*Alternative Annotator Test* proposé par (Calderon *et al.*, 2025). Ce test permet de déterminer, pour une tâche et un critère donnés, si les annotations produites par un modèle peuvent se substituer à celles d'un annotateur humain avec un niveau de fiabilité suffisant. Une fois calibré, le *LLM-as-Judge* sera appliqué à de nouveaux exercices, indices et feedbacks générés, non comme instance de validation finale, mais comme outil de pré-filtrage. Les annotations humaines jouent ainsi un double rôle : elles documentent la qualité

des productions générées et fournissent un signal de référence pour calibrer, limiter et justifier l'usage du *LLM-as-Judge* à grande échelle.

Le projet comprend également un important volet de science ouverte. Une mise à disposition sur Hugging Face est en cours de préparation, incluant des traces d'apprentissage d'élèves collectées en conditions réelles sur Adaptiv'Math et MIA Seconde, les contenus des exercices correspondants, les métadonnées pédagogiques associées, ainsi que les graphes de prérequis qui décrivent l'organisation des activités et les relations de dépendance pédagogique entre elles. Ces ressources seront accompagnées de scripts de prétraitement, de code d'évaluation et d'un outil de visualisation permettant l'inspection des trajectoires d'apprentissage, des usages et de la structure des parcours pédagogiques. Les données diffusables font l'objet d'un processus d'anonymisation et de minimisation : les identifiants des élèves, classes et sessions sont anonymisés, les informations temporelles sont exprimées relativement à une référence temporelle anonymisée, et les noms, informations démographiques libres, noms d'établissements ou identifiants géographiques ne sont pas diffusés. Ce volet vise à favoriser, dans un cadre respectueux de la protection des élèves, la reproductibilité des analyses, la comparaison des approches de modélisation et de génération, ainsi que le développement d'outils méthodologiques partagés dans le champ de l'IA en éducation.

3 Perspectives d'évaluation et conclusion

À ce stade, GAIMHE doit donc être compris comme une contribution de cadrage méthodologique et d'ingénierie pédagogique, dont les résultats empiriques seront établis lors des déploiements et évaluations contrôlées à venir. À court terme, GAIMHE vise à finaliser la calibration du *LLM-as-Judge*, à consolider la chaîne de validation hybride et à poursuivre l'ouverture des ressources produites. À plus long terme, l'enjeu central sera l'évaluation des effets des contenus générés en contexte scolaire réel. La question principale est de déterminer si leur intégration dans le STI permet d'enrichir le système sans dégrader la qualité des apprentissages.

L'évaluation en contexte réel prendra la forme d'un déploiement contrôlé en classe, prioritairement sur des activités de mathématiques, physique et/ou de littératie, du primaire au lycée, selon les plateformes et les disciplines retenues. Le protocole visera à comparer au moins deux conditions : une condition STI classique, utilisant le séquençage adaptatif existant avec des contenus experts, et une condition hybride intégrant des contenus ou feedbacks générés puis validés. Selon les contraintes de terrain, une condition contrôle non adaptative ou "pratique habituelle" pourra également être intégrée. Les variables dépendantes incluront les gains d'apprentissage pré/post-test, la progression dans les objectifs du graphe pédagogique, la persistance dans les activités, le temps passé, les erreurs récurrentes et des indicateurs de motivation.

GAIMHE défend ainsi une approche de l'IA en éducation qui ne repose ni sur le rejet des modèles génératifs, ni sur leur intégration sans garde-fous, mais sur une articulation raisonnée entre génération, validation humaine, données réelles d'usage et pilotage adaptatif. Au-delà du développement technologique, GAIMHE ambitionne également de contribuer à la structuration méthodologique du domaine en adoptant une approche de science ouverte, par la diffusion de corpus, de graphes, de traces et d'outils d'analyse. En ce sens, il constitue à la fois un projet d'ingénierie pédagogique, un cadre de recherche sur les usages responsables de l'IA générative, et une proposition pour renforcer la transparence et le caractère cumulatif de la recherche en IA pour l'éducation (Baker *et al.*, 2024; Topham *et al.*, 2025).

Références

ABDELRAHMAN G., WANG Q. & NUNES B. P. (2023). Knowledge tracing : A survey. *ACM Computing Surveys*, **55**(11), 1–37. DOI : 10.1145/3569576.

ANDERSON J. R., CORBETT A. T., KOEDINGER K. R. & PELLETIER R. (1995). Cognitive tutors : Lessons learned. *Journal of the Learning Sciences*, **4**(2), 167–207. DOI : 10.1207/s15327809jls0402_2.

BAKER R. S., HUTT S., BROOKS C. A., SRIVASTAVA N. & MILLS C. (2024). Open science and educational data mining : Which practices matter most ? In *Proceedings of the 17th International Conference on Educational Data Mining*, p. 279–287.

BASTANI H., BASTANI O., SUNGU A., GE H., KABAKCI Ö. & MARIMAN R. (2025). Generative ai without guardrails can harm learning : Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, **122**(26), e2422633122.

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).

CLÉMENT B., SAUZÉON H., ROY D. & OUDEYER P.-Y. (2024). Improved performances and motivation in intelligent tutoring systems : Combining machine learning and learner choice. arXiv :2402.01669.

COLLIOT T., KRICHEN O., GIRARD N., ANQUETIL É. & JAMET É. (2024). What makes tablet-based learning effective ? a study of the role of real-time adaptive feedback. *British Journal of Educational Technology*, **55**(5), 2278–2295.

HAYES A. F. & KRIPPENDORFF K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, **1**(1), 77–89. DOI : 10.1080/19312450709336664.

JURENKA I., KUNESCH M., MCKEE K. R., GILLICK D., ZHU S., WILTBERGER S., PHAL S. M., HERMANN K., KASENBERG D., BHOOPCHAND A. ET AL. (2024). Towards responsible development of generative ai for education : An evaluation-driven approach. *arXiv preprint arXiv :2407.12687*.

KASNECI E., SESSLER K., KÜCHEMANN S., BANNERT M., DEMENTIEVA D., FISCHER F. & KASNECI G. (2023). Chatgpt for good ? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, **103**, 102274. DOI : 10.1016/j.lindif.2023.102274.

TOPHAM L., ATHERTON P., REYNOLDS T., HUSSAIN Y., HUSSAIN A., KOLIVAND H. & KHAN W. (2025). Artificial intelligence in educational technology : A systematic review of datasets and applications. *ACM Computing Surveys*, **58**(3), 1–28.

TRICOMI E. & DEPASQUE S. (2016). The role of feedback in learning and motivation. In *Recent developments in neuroscience research on human motivation*, volume 19, p. 175–202. Emerald Group Publishing Limited.