Résumé scientifique du projet

Un tuteur d'intelligence artificielle souverain, hors-ligne et socratique pour l'autonomie cognitive et la souveraineté numérique européenne

Lucas Gerhardt Clément Bellet-Odent
Lycée Henri Matisse, Vence (France) · Encadrant : P. Aboussouan · EUCYS 2026, Kiel · Domaine : Computing

---

Le problème. En France, 60 % des élèves déclarent mal organiser leurs révisions, et les outils d'IA générative (ChatGPT, Photomath, Gemini) fournissent des réponses immédiates qui suppriment l'effort de récupération pourtant nécessaire à l'apprentissage. Une étude par électroencéphalographie du MIT (2025) mesure chez les utilisateurs de ces outils une chute marquée de la connectivité cérébrale, un phénomène que les chercheurs nomment « dette cognitive ». À l'inverse, une étude de Harvard montre qu'une IA configurée en tuteur, brève et sans révéler la solution, double les gains d'apprentissage par rapport à un cours actif. Le facteur déterminant n'est donc pas la puissance du modèle, mais sa *posture*.

Notre solution. Nous présentons un assistant scolaire fondé sur l'intelligence artificielle qui s'exécute **entièrement sur l'appareil de l'élève**, sans aucune connexion. Plutôt que de livrer la réponse, il adopte une **posture socratique** stricte : il ne donne jamais la solution, mais guide l'élève par le questionnement, un pas à la fois. À partir des cours réels de l'élève, il génère quiz, fiches de révision et séances de tutorat, tout en conservant les données d'apprentissage (cours, conversations, notes) sur l'appareil.

La technologie. L'assistant repose sur les petits modèles ouverts **Gemma** de Google, *quantifiés par conception* en int4 par *quantization-aware training* (QAT) : la quantification compresse le modèle pour qu'il tienne sur un téléphone (un modèle de 4 milliards de paramètres passe de 8 Go à environ 2,6 Go), et en intégrant cette compression au processus d'entraînement, la QAT préserve une qualité quasi identique à la pleine précision. L'inférence s'exécute sur le processeur graphique du téléphone via un moteur **GGUF**. Un *fine-tuning* **QLoRA**, mené sur un corpus socratique de 1 000 paires validées par des enseignants, instille le comportement maïeutique en n'entraînant que deux petites matrices de rang faible (environ 0,1 % des paramètres), ce qui rend la spécialisation légère et réversible. Pour la fiabilité, une **recherche augmentée (RAG)** entièrement embarquée combine une recherche lexicale (BM25) et une recherche sémantique (EmbeddingGemma), fusionnées par *Reciprocal Rank Fusion*, afin d'ancrer chaque réponse dans les manuels réels de l'élève et de réduire les hallucinations. Une **grammaire de décodage** (GBNF) force la structure des « coups » socratiques et rend la réponse littérale grammaticalement inatteignable, déplaçant la posture pédagogique de la couche du *prompt* (contournable) vers celle du *décodage* (structurelle). Enfin, un **modèle de données canonique** projette chaque environnement numérique de travail (ENT) sur une couche commune via un adaptateur unique, au-dessus de laquelle toute l'intelligence pédagogique est écrite une seule fois ; plusieurs connecteurs sont déjà implémentés (Pronote, ClasseViva, WebUntis, Moodle, Canvas), rendant l'outil portable à l'échelle des **37,5 millions d'élèves** du secondaire de l'Union européenne. Six contributions de recherche originales (PAQ, PEDAL, ZED/W-Gap, ZPD-Steer, CDL/SRR, PS2) prolongent cette architecture en imposant la rétention de la réponse comme invariant de *décodage*, de *quantification* et de *récupération*, et en estimant l'état cognitif de l'élève entièrement sur l'appareil.

Souveraineté et confidentialité. Parce que l'inférence et les données d'apprentissage ne quittent pas l'appareil, l'architecture répond *par construction* au RGPD et à l'AI Act, et sert la souveraineté numérique européenne : aucun datacenter à bâtir, aucun contenu scolaire transféré vers un cloud extra-européen. Seule une progression non nominative (XP, niveau) est synchronisée pour alimenter le classement de la gamification.

Résultats et impact. En évaluation à l'aveugle (40 questions, 2 correcteurs), le *fine-tuning* installe une posture socratique nette (**4,6/5**), absente de ChatGPT (1,6/5, qui livre la réponse) comme du modèle de base non spécialisé (Llama 3.2, 1,2/5) : c'est la spécialisation, non la puissance du modèle, qui crée la posture. **87 %** des élèves interrogés font confiance à l'application pour leurs données, contre 19 % à ChatGPT. L'inférence est **frugale** (environ 3 W, l'ordre de grandeur d'une ampoule LED), sans transit cloud ni eau de refroidissement. Le projet agit ainsi sur trois fractures européennes : l'*équité* (un tuteur gratuit et hors-ligne, sur un appareil modeste, qui atteint les élèves que le cloud abandonne), l'*écologie* (une IA frugale qui prolonge la durée de vie des appareils anciens) et la *souveraineté* (une IA qui appartient à l'élève et reste en Europe). *La meilleure IA éducative n'est pas la plus puissante : c'est celle qui sait quand ne pas répondre.*