European Union Contest for Young Scientists

EUCYS 2026 · Kiel, Allemagne · Domaine : Computing

Un tuteur d'IA souverain, hors-ligne et socratique

pour l'autonomie cognitive et la souveraineté numérique européenne

Lucas Gerhardt Clément Bellet-Odent

Lycée Henri Matisse, Vence (France) · Professeur encadrant : Pierre Aboussouan

Rapport scientifique, délégation française

Notre assistant est un outil scolaire fondé sur l'intelligence artificielle qui s'exécute **entièrement sur l'appareil de l'élève**. Là où les outils génératifs livrent la réponse, nourrissant une « **dette cognitive** » mesurée par électroencéphalographie, notre assistant adopte une **posture socratique** stricte : il ne donne jamais la solution, mais guide par le questionnement. Propulsé par de petits modèles ouverts (**Gemma**, quantifiés *par conception* en int4 QAT), spécialisés par *fine-tuning* **QLoRA** et ancrés par une **recherche augmentée embarquée** (RAG), il transforme les cours réels de l'élève en quiz, fiches et tutorat, **sans que ses cours ni ses conversations ne quittent le téléphone**. Ce rapport établit, mesures à l'appui, qu'une IA éducative *privée, frugale et souveraine* peut, sur le critère qui compte le plus pour apprendre (**faire raisonner l'élève**), faire mieux que les modèles cloud.

PARTENAIRES, SOUTIENS & COLLABORATIONS

© 2026, Lycée Henri Matisse, Vence, France · 23 juin 2026, 19 :38 UTC (GMT+0)

Un tuteur d'IA 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

Notre assistant est **en ligne sur l'App Store**, gratuit et hors-ligne, adaptable à tout curriculum et toute langue. La vidéo de présentation (2 min) est jointe au dossier transmis ; ce rapport en constitue le supplément et l'approfondissement scientifique.

Résumé — L'IA générative répond à la place de l'élève et court-circuite l'effort cognitif, une étude EEG du MIT (2025) y associe un effondrement de la connectivité cérébrale (« dette cognitive »). Notre assistant prend le problème à rebours : un tuteur *socratique* qui ne donne jamais la réponse mais guide par le questionnement, exécuté **100 % sur l'appareil**. Il repose sur de petits modèles Gemma **quantifiés par conception** (int4 *QAT*, moteur GGUF), spécialisés par *fine-tuning* QLoRA sur un corpus socratique validé par des enseignants (1 000 paires, licence libre) et ancrés par un RAG hybride embarqué. Les données d'apprentissage de l'élève (cours, conversations, notes) restent sur l'appareil, seule la progression non nominative (XP, niveau) alimente le classement : l'architecture *répond par construction* au RGPD et à l'AI Act et sert la souveraineté numérique européenne. Un modèle de données canonique projette **14 ENT** (Pronote, ClasseViva, WebUntis, Moodle…) sur une couche pédagogique mutualisée, rendant l'outil portable aux **37,5 M d'élèves du secondaire de l'UE** (∼87 M avec les États-Unis via Google Classroom). 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, à ∼3 W et sans transit cloud. Six contributions originales (**PAQ**, **PEDAL**, **ZED/W-Gap**, **ZPD-Steer**, **CDL/SRR**, **PS2**) déplacent la rétention de la réponse du *prompt* vers le *décodage*, la *quantification* et la *récupération*, chacune rapportée selon un statut honnête, sans résultat fabriqué. *La meilleure IA éducative n'est pas la plus puissante : c'est celle qui sait quand ne pas répondre.*

Mots-clés — IA embarquée ; tutorat socratique ; souveraineté numérique européenne ; RGPD *by-design* ; quantification & QLoRA ; RAG ; frugalité.

---

1 Introduction

En France, 60 % des élèves déclarent mal organiser leurs révisions (PISA 2022), et les outils génératifs (ChatGPT, Photomath, Gemini) fournissent des réponses immédiates qui suppriment l'effort de récupération nécessaire à l'apprentissage. L'étude EEG *Your Brain on ChatGPT* du MIT Media Lab [9] ($N = 54$, 4 sessions) mesure chez les utilisateurs de LLM la **connectivité cérébrale la plus faible** (−55 % vs. le groupe sans outil) et observe que **83 %** d'entre eux sont incapables de citer une phrase qu'ils venaient d'écrire (contre ∼11 %). Les chercheurs nomment ce phénomène la « **dette cognitive** ». À l'échelle de l'Union, 37,5 M de collégiens et lycéens [1] y sont exposés.

L'inverse est pourtant possible. Une étude de Harvard [7] ($N = 194$) montre qu'une IA configurée en **tuteur** (brève, « un pas à la fois », sans révéler la solution) double les gains d'apprentissage par rapport à un cours actif. *Le facteur déterminant n'est pas la puissance du modèle, mais sa posture.*

1. Eurostat, *Secondary education statistics*, 2024 (CITE 2+3).

Cet enjeu est aussi devenu européen et réglementaire. L'AI Act classe l'éducation « à haut risque » (obligations applicables dès août 2026), le RGPD impose la minimisation des données des mineurs, et le *Tech Sovereignty Package* de la Commission (2026) rappelle que l'UE dépend de pays non-européens pour plus de **80 %** de ses technologies numériques clés. Une IA éducative traitant les données *localement* répond à ces trois exigences (pédagogique, juridique et souveraine) d'un seul geste architectural.

Problématique. *Comment concevoir un assistant scolaire fondé sur l'IA qui renforce l'autonomie cognitive de l'élève au lieu de s'y substituer, tout en garantissant la confidentialité des données des mineurs et la souveraineté numérique européenne ?*

Contributions. (i) Un tuteur **socratique** contraint par *fine-tuning* et *system prompt* ; (ii) une architecture **Local-First** où toute l'inférence est embarquée, incarnant le RGPD et l'AI Act *by-design* ; (iii) un **RAG hybride embarqué** (lexical + sémantique) réduisant les hallucinations par ancrage sur des sources locales ; (iv) un **modèle de données canonique** rendant l'outil portable à 14 systèmes scolaires européens ; (v) une **évaluation à l'aveugle** montrant que la posture socratique vient de la spécialisation (notre *fine-tuning*), non de la puissance du modèle (ni ChatGPT ni le Llama brut ne l'ont) ; et surtout (vi) **six contributions originales** (§3, Tab. 2), dont **PAQ**, une quantification int4 pilotée par la posture, *piste de recherche non encore mesurée*, qui déplacent la posture socratique du *prompt* vers le *décodage*, la *quantification*, la *récupération* et la *mesure*, entièrement sur l'appareil.

2 Contexte scientifique

Six corpus fondent notre conception. La *théorie de la charge cognitive* [1] impose de minimiser la charge extrinsèque et de guider « un pas à la fois ». La *répétition espacée* [2, 3] (méta-analyse : 839 mesures) et l'*effet de test* [4] justifient flashcards et quiz issus du cours réel. La *théorie de l'autodétermination* [5] fonde une gamification de l'effort. L'*approche socratique assistée* [6, 7] montre qu'une IA contrainte à questionner apprend mieux. Enfin, l'*Edge AI* [11] rend possible un traitement local des données sensibles. Le paradoxe de l'apprentissage actif [8] (on apprend plus mais on *ressent* apprendre moins) justifie d'assumer la friction socratique.

Ces résultats convergent vers une exigence quantifiable : réduire la charge extrinsèque (guidage progressif), forcer la récupération active (quiz, flashcards à intervalles croissants), soutenir la motivation intrinsèque (gamification de l'effort), et traiter les données localement.

Lacune. À notre connaissance, aucune solution n'associe *simultanément* : posture socratique stricte, confidentialité 100 % locale, gamification de l'effort, et portabilité/souveraineté paneuropéennes. ChatGPT et Photomath livrent la réponse (cloud, non souverains) ; les tuteurs cloud type Khanmigo restent dépendants d'une connexion et d'un hébergement tiers ; les modèles ouverts bruts exécutés localement n'ont ni posture pédagogique ni ancrage curriculaire. Notre assistant occupe seul l'intersection (Tab. 1).

Mise en garde. Ce tableau est, par construction, partiellement subjectif : d'autres grilles avantageraient chaque concurrent (ChatGPT sur la qualité brute, Khanmigo sur l'écosystème). Notre assistant ne prétend pas être « meilleur partout » ; il revendique d'occuper *seul* l'intersection socratique *et* 100 % local *et* souverain.

3 Contributions originales

Au-delà de son architecture souveraine et de sa posture maïeutique mesurée, notre assistant introduit des contributions qui déplacent la pédagogie socratique de la couche du *prompt*, fragile et contournable, vers le **décodage**, la **récupération** et la **mesure**, en régime int4/GGUF sur matériel grand public. Chacune est une *composition défendable de primitives établies*, étiquetée

honnêtement, implémenté, prototype ou proposé (protocole pré-enregistré) : aucun chiffre n'est avancé sans protocole non fabricable. Toutes partagent un unique **état-apprenant local** : *la rigueur pédagogique est imposée par construction, sur le téléphone, non sollicitée par instruction.*

Thèse : la posture socratique *par construction*, au décodage

Un *system prompt* maïeutique se contourne par ingénierie sociale. Nos contributions rendent la posture **structurelle** : la grammaire de décodage rend le bloc-réponse *grammaticalement inatteignable* (PEDAL, §5.6) ; la réponse d'or locale devient une *contrainte négative* jamais exfiltrée (ZED, §5.7) ; l'état cognitif de l'élève, estimé sur l'appareil, pilote récupération, génération et étayage (ZPD-Steer, CDL, PS2) ; et **PAQ** (§5.3) alloue jusqu'aux *bits de quantification* pour préserver la non-réponse. La pédagogie n'est plus une consigne : c'est une propriété du système, jusque dans ses poids.

4 Hypothèses et protocole

Nous traitons notre assistant comme objet de recherche via six hypothèses testables (Tab. 4) : posture socratique (**H1**), confiance par le local (**H2**), fonctionnement hors-ligne (**H3**), orchestration locale/cloud (**H4**), portabilité par modèle canonique (**H5**), frugalité énergétique (**H6**). La vérification combine une **évaluation à l'aveugle** (40 questions, 5 matières, 2 correcteurs), un **questionnaire** ($N = 47$ lycéens, échelle de Likert), des **benchmarks** sur 5 appareils, et un **audit d'architecture**.

5 Système : une IA éducative locale et souveraine

5.1 Architecture *Local-First*

Notre premier prototype (V1) reposait sur l'API Gemini : toutes les conversations transitaient par des serveurs tiers, compromettant à la fois la confidentialité (**H2**) et l'usage hors-ligne (**H3**). Ce constat (un assistant « privé » qui envoyait pourtant chaque mot d'un mineur vers un cloud extra-européen) a déclenché le pivot vers l'embarqué. Le principe directeur (Fig. 5) devient alors : les données scolaires sont récupérées *directement* depuis le serveur de l'établissement et restent sur l'appareil ; l'inférence IA s'y exécute intégralement ; aucun serveur intermédiaire n'est interposé. Un orchestrateur cloud/local (**H4**) a été prototypé (bascule <800 ms en laboratoire) puis mis en pause face à des limites d'infrastructure, au profit d'un fonctionnement **100 % local**, ce qui, paradoxalement, a *renforcé* la thèse de souveraineté.

Le choix du modèle embarqué résulte d'une évaluation comparative en conditions réelles (iPhone 14, Samsung S23). Llama 3 (8B) a été écarté (>15 s/réponse, >4 Go de RAM, expérience inacceptable) au profit d'un modèle 4B (latence médiane 2,7 s), divisant le temps de réponse par 5 ; un mécanisme de chargement/déchargement dynamique combiné à la quantification 4 bits a ramené l'empreinte mémoire sous le seuil des appareils à moins de 4 Go.

5.2 Modèle embarqué et inférence

Le modèle s'exécute sur le GPU du téléphone via un moteur **GGUF** (`llama.rn`/`llama.cpp`) : contexte 4 096 tokens, cache KV quantifié, *flash attention*, et une **génération JSON contrainte par grammaire** (GBNF) qui rend quiz et fiches structurellement valides. Nous adoptons les modèles ouverts **Gemma** en **int4 *quantization-aware training*** (QAT) [12], Gemma 3, et désormais **Gemma 4** (juin 2026). La QAT **intègre la quantification *à* l'entraînement** (cible : le modèle pleine précision), si bien que le 4 bits conserve une qualité *quasi identique* au bf16 : le 4B passe de **8 Go (bf16) à ∼2,6 Go en int4**, la dégradation de perplexité est **réduite d'environ 54 %** face à une quantification post-entraînement, et le plus petit Gemma 4 (E2B)

tient en **moins de 1 Go**. Un modèle **quantifié *par conception***, meilleure qualité *on-device* à empreinte moindre (→**H3**, **H6**).

5.3 PAQ, quantification sensible à la posture (*contribution phare*, H1, H6)

Hypothèse (à tester). Nous faisons l'hypothèse que la quantification int4 **éroderait silencieusement la posture socratique** : sous compression, le modèle régresserait vers le réflexe de *donner la réponse*, alors même que la **perplexité ne bougerait presque pas**, un effet documenté pour l'alignement en général, mais jamais étudié sous l'angle pédagogique. *Mécanisme.* L'allocateur de `llama.cpp` (matrice d'importance, lignée OBQ/GPTQ) pondère chaque colonne de poids par son **énergie d'activation** (moyenne des carrés des activations) sur un corpus générique, il préserve donc ce que le modèle fait *le plus souvent* : répondre. PAQ remplacerait ce critère par une **marge comportementale contrastive** *socratique-vs-réponse*, dérivée de notre métrique de fuite (W-Gap, §5.7), et **routerait le budget de bits int4 vers les poids qui portent la non-réponse**. À partir de deux jeux d'invites construits sur un sous-ensemble tenu à l'écart de notre corpus socratique validé, P_{soc} (comportement socratique) et son contrefactuel P_{ans} (« donne directement la réponse »), on calcule, par colonne i, l'énergie d'activation $A_{\text{soc},i}$ et une saillance comportementale g_i, puis un score d'importance-posture :

$$s_i = A_{\text{soc},i}\Big(1 + \lambda \,\frac{|g_i|}{\operatorname{med}_j |g_j|}\Big), \quad g_i = \mathbb{E}\big[a_i \,\partial_{W_i}\big(\log p(t_{\text{soc}}) - \log p(t_{\text{ans}})\big)\big]. \tag{1}$$

Injectée dans la matrice d'importance, s_i réoriente l'erreur de quantification int4 vers les directions porteuses de posture, *à type de tenseur et budget de bits inchangés* : on obtiendrait un GGUF int4 de **taille et latence identiques**, dont l'erreur serait minimisée le long des directions qui préservent la non-réponse.

Nouveauté. À notre connaissance, première allocation de bits int4 pilotée par une *marge comportementale contrastive*, formulant la non-réponse socratique comme *posture fragile à la quantification*, distincte d'AWQ (magnitude d'activation seule). **Statut.** Prototype ; aucun chiffre avant exécution. **Garde-fou *(pré-enregistré).*** Si la posture est *diffuse* (s_i non concentré), PAQ se réduit à la matrice d'importance standard et est rapporté *sans gain revendiqué*. **Validation.** Trois quantifications à budget égal, B0 (sans matrice), B1 (générique), B2 (PAQ), comparées au fp16 sur banc tenu à l'écart (fuite W-Gap, posture, perplexité, tok/s, énergie). Critère : *à taille et latence égales, PAQ récupère-t-il vers le fp16 une part de la posture que l'int4 standard perd ?*

5.4 Spécialisation : dataset et *fine-tuning* QLoRA

Les modèles généralistes ignorant la terminologie des programmes, nous les spécialisons sur un corpus socratique. Une amorce de ∼1 200 paires rédigées par des enseignants alimente une **distillation de connaissances** (plateforme interne *Lucid Labs*) ; après **validation enseignante** (interface *Lucid Professeur*), **1 000 paires** retenues sur ∼10 000 candidats générés (938 « gold », soit 93,8 %) couvrant 18 matières et 7 niveaux (6e–Tle) constituent le corpus, **publié sous licence libre** (CC-BY-NC-SA 4.0). Le *fine-tuning* emploie QLoRA [13] (LoRA [14] sur modèle gelé 4 bits NF4) : le modèle de base W_0 reste *figé* et l'on n'entraîne que deux petites matrices de rang faible, A (compression) et B (décompression), Fig. 2. Chaque couche calcule alors

$$h \;=\; W_0\, x \;+\; \frac{\alpha}{r}\,\underbrace{B\,A}_{\Delta W}\, x, \qquad B \in \mathbb{R}^{d\times r},\; A \in \mathbb{R}^{r\times k},\; r \ll \min(d,k). \tag{2}$$

Avec $r = 16$, seuls ∼0,1 % des paramètres sont modifiés ($r = 4$ instabilise la posture, $r = 64$ sur-apprend) ; la perte de validation finale $\mathcal{L} \approx 0{,}87$ se stabilise sans divergence train/val (pas de sur-apprentissage manifeste). Un *system prompt* maïeutique strict (ne jamais donner la réponse, identifier le blocage, questionner par paliers) renforce l'effet du *fine-tuning* (→**H1**),

comme l'illustre le dialogue de la Fig. 3.

5.5 Fiabilité : un RAG hybride embarqué

Le *fine-tuning* seul laissait des erreurs factuelles. Nous ajoutons une **recherche augmentée (RAG)** [15] *entièrement embarquée* (Fig. 6), plus robuste qu'une simple recherche vectorielle. Les manuels sont découpés en *chunks* bornés par phrases (∼900 caractères). À chaque question, deux recherches s'exécutent sur l'appareil, **lexicale BM25** (k_1=1,2, b=0,75, titres ×3) et **sémantique** via **EmbeddingGemma** (vecteurs 256-dim, troncature Matryoshka, normalisation L2), puis sont fusionnées par ***Reciprocal Rank Fusion*** ($1/(60 + \text{rang})$). Les meilleurs passages sont injectés dans le contexte : les réponses deviennent **précises, locales et sourcées**.

5.6 PEDAL, la rétention de la réponse comme invariant de décodage (H1)

PEDAL transforme la posture maïeutique « ne jamais donner la réponse » d'une requête de *prompt*, contournable par ingénierie sociale, en un **invariant de décodage**. *Couche 1 (structurelle)* : nous compilons une taxonomie de coups pédagogiques (diagnostic, indice optionnel, question obligatoire) en une grammaire GBNF passée à `llama.cpp`/Gemma via le champ `grammar` ; aucune production ne permet de terminer un tour sur un bloc-solution, rendant le modèle *structurellement* incapable de conclure sur une réponse nue. Un automate fini (nombre de tentatives, palier d'indice, persisté dans WatermelonDB) sélectionne la variante de grammaire à chaque tour, ce qui contre aussi l'effondrement de posture multi-tours. *Couche 2 (lexicale)* : pour les problèmes dont la réponse d'or est disponible localement (quiz, flashcards, curriculum ancré par RAG), nous tokenisons la réponse canonique et ses variantes de surface, puis les injectons en interdiction dure (`logit_bias`), activable *paresseusement* après l'état diagnostic (`grammar_lazy`/`grammar_triggers`) pour laisser l'étayage initial libre. Formellement, à l'étape t du décodage, les logits $\ell_t \in \mathbb{R}^{|V|}$ sont corrigés selon l'état s_t de l'automate :

$$\tilde{\ell}_{t,i} = \ell_{t,i} + \underbrace{\log \mathbf{1}[i \in \mathcal{A}(s_t)]}_{\text{couche 1 : grammaire}} - \underbrace{\beta\, \mathbf{1}[i \in \mathcal{B}]}_{\text{couche 2 : ban réponse}}, \quad \beta \to +\infty, \tag{3}$$

où $\mathcal{A}(s_t)$ est l'ensemble des tokens admis par la grammaire GBNF dans l'état s_t (le terme vaut 0 si admis, $-\infty$ sinon) et $\mathcal{B}$ l'ensemble des tokens des formes de surface de la réponse d'or locale : le bloc-réponse devient de probabilité *nulle par construction*. À notre connaissance, c'est le premier système à imposer *simultanément* la structure des coups socratiques et la rétention littérale de la réponse comme contrainte de décodage sur un LLM quantifié de téléphone.
Nouveauté. Composition inédite de primitives établies (interdictions lexicales à la NeuroLogic [19], décodage sous grammaire GBNF, taxonomies d'actes pédagogiques). **Statut.** Prototype (Couche 1 immédiate ; Couche 2 sur le sous-ensemble à réponse connue). **Validation.** Étude A/B/C pré-enregistrée (prompt seul / +grammaire / +grammaire+ban) sur problèmes tenus à l'écart et attaques adverses ; métriques : fuite, posture (grille aveugle, κ de Cohen), *taxe d'alignement* (justesse du premier coup, tok/s, watts). Un résultat négatif reste publiable.

5.7 ZED, garde de décodage à égression zéro & métrique W-Gap (H1, H2)

ZED requalifie la rétention de réponse en invariant de décodage *et* fournit sa métrique. *Niveau 1 (implémenté, garanti)* : la réponse d'or étant locale (chunk RAG, curriculum, solution générée), nous tokenisons ses formes de surface et les bannissons en dur (`logit_bias`, confirmé dans notre build `llama.rn`), rendant la séquence-réponse littérale *non émettable* quel que soit le *prompt*, garantie de correction robuste au *jailbreak*, non un score empirique. *Niveau 2 (prototype)* : pour les fuites par *paraphrase* que l'interdiction lexicale manque, un discriminateur inspiré de FUDGE [18] réutilise EmbeddingGemma (256-dim) pour scorer la proximité de la réponse partielle à la

réponse d'or, par génération à fenêtres redémarrées (`llama.rn` n'exposant aucun crochet logit par pas, limitation déclarée, coût de re-préfill mesuré). Le prochain token est rééchantillonné selon

$$P(x_t \mid x_{<t}) \;\propto\; P_{\mathrm{LM}}(x_t \mid x_{<t}) \cdot P_\phi(\text{non-fuite} \mid x_{\le t})^\eta \,, \tag{4}$$

où P_ϕ est le discriminateur logistique embarqué sur EmbeddingGemma et $\eta > 0$ la force de guidage. La métrique de fuite **W-Gap**, déterministe et locale, agrège recouvrement et sémantique *sans juge cloud* :

$$\text{W-Gap}(\hat{y}, y^\star) = \lambda_1 \frac{|g_n(\hat{y}) \cap g_n(y^\star)|}{|g_n(y^\star)|} + \lambda_2 \, \mathbf{1}[\text{match num.}] + \lambda_3 \cos\big(e(\hat{y}), e(y^\star)\big), \tag{5}$$

avec g_n les n-grammes, $e(\cdot)$ EmbeddingGemma et $\sum_k \lambda_k = 1$; une *fuite dure* est déclarée si la forme finale exacte ou $\ge k$ tokens d'or contigus apparaissent.
Nouveauté. Premier garde de rétention de réponse au décodage *et* première métrique de fuite (W-Gap) fonctionnant entièrement sur appareil (régime int4/GGUF, sans juge cloud) pour le tutorat.
Statut. Niveau 1 : implémenté & mesurable (garantie de correction) ; Niveau 2 et validation de W-Gap : prototype/préliminaire. **Validation.** Protocole d'élève adverse sur trois bras ; le résultat-vedette honnête est la réduction de W-Gap sur les fuites *paraphrasées* (le 0 % de fuite littérale du Niveau 1 est un contrôle de cohérence, pas une découverte). Validité de W-Gap : corrélation (Spearman/Pearson) avec les jugements humains sur 40 questions, donc *préliminaire* ($N = 40$).

5.8 ZPD-Steer, l'état-apprenant local pilote récupération et génération (H1, H2)

Le moteur de révision de notre assistant (`services/vacation/elo.ts`) maintient déjà une estimation d'habileté *de type Elo*, formellement apparentée au modèle de Rasch [20], habileté θ par matière, difficulté λ par notion, mise à jour en ligne depuis le journal de révision local ; mais ce signal est aujourd'hui cloisonné dans la répétition espacée et invisible au tuteur, qui n'adapte la difficulté que via une chaîne statique de niveau de classe. ZPD-Steer relie cet état-apprenant à la boucle socratique de deux manières, *entièrement sur l'appareil*. (1) **ZPD-Rerank** : un score de difficulté/lisibilité est précalculé par chunk à l'indexation, puis un terme à coût quasi nul $\lambda \cdot f(d_{\text{chunk}} - \theta)$ est fusionné dans le bloc *Reciprocal Rank Fusion* existant, favorisant les passages *légèrement* au-dessus de la maîtrise (difficulté désirable au sens de Bjork [22], zone proximale de Vygotsky [21]). (2) **ZPD-Decode** : θ, λ et les notions récemment fragiles sont injectés comme état structuré dans le *prompt* socratique, et la fonction `targetDifficulty`, déjà implémentée, choisit la bande de difficulté de la prochaine question. La personnalisation vit dans l'*espace mémoire local*, non dans les poids : sans oubli catastrophique, et instantanément effaçable. L'habileté θ provient d'un modèle de Rasch, la probabilité qu'un élève d'habileté θ réussisse un item de difficulté b est

$$P(\text{réussite} \mid \theta, b) = \sigma(\theta - b), \qquad \sigma(z) = \tfrac{1}{1+e^{-z}}, \tag{6}$$

θ étant mis à jour en ligne depuis le journal local, et le reclassement d'un chunk c pour la requête q devient

$$\text{score}(c \mid q, \theta) = \underbrace{\sum_{r \in \{\text{BM25},\, \text{sém.}\}} \frac{1}{60 + \text{rang}_r(c)}}_{\text{RRF (existant)}} + \lambda \, g(d_c - \theta) \,, \tag{7}$$

où d_c est la difficulté du chunk et g un noyau de *difficulté désirable* maximal en un faible décalage positif $\delta^\star$ (p. ex. $g(u) = -(u - \delta^\star)^2$) : on favorise les passages *juste* au-dessus de θ.
Nouveauté. À notre connaissance, première mise en œuvre *embarquée* reclassant une récupération de corpus en texte libre selon l'ajustement entre difficulté du passage et habileté estimée d'un *apprenant humain*, les tuteurs adaptatifs et le *knowledge tracing* estiment l'habileté de longue date,

mais côté cloud et hors récupération RAG en texte libre ; ici l'estimation est calculée 100 % sur l'appareil. **Statut.** ZPD-Rerank : prototype (mécanisme implémenté, gain à mesurer) ; boucle fermée de décodage : prototype/préliminaire. **Validation.** A/B aveugle pré-enregistré (ZPD-Rerank *vs* pertinence seule) sur ∼150 tuples (requête × niveau simulé), métrique = appropriation de niveau ; les étiquettes de difficulté sont validées contre le jugement enseignant *avant* l'A/B. Jeu publié sous licence libre.

5.9 PS2, ordonnanceur de lutte productive par commutation de grammaire (H1)

PS2 porte l'idée de *wheel-spinning* et de lutte productive, issue des tuteurs intelligents structurés par étapes [23], dans le chat socratique en texte libre. À partir de la seule trace locale (tentatives avant succès, temps par étape, séquences erreur-puis-autocorrection, erreurs répétées-identiques *vs* variées mesurées par similarité EmbeddingGemma), un classifieur estime une **bande de lutte à trois états** qui *commute la variante de grammaire de coups* au décodage : en lutte productive, la grammaire interdit les indices et force un terminal-question ; en frustration, elle autorise un palier d'indice (voire un exemple travaillé d'un *autre* problème, sûr en fuite) ; en sous-stimulation, elle saute l'étayage. La bande β est choisie par un classifieur P_ψ sur le vecteur de signaux $\mathbf{z}$ (tentatives, temps par étape, similarité inter-erreurs) et indexe la variante de grammaire $\mathcal{A}^{(\beta)}$ de l'éq. (3) :

$$\beta = \underset{k \in \{\text{sous-stim., prod., frust.}\}}{\arg\max} P_\psi(k \mid \mathbf{z}) \longmapsto \mathcal{A}^{(\beta)}(s_t). \tag{8}$$

La commutation utilise `grammar_lazy`/`grammar_triggers`. PS2 opérationnalise ainsi la difficulté désirable dans le chat LLM, qui n'a aujourd'hui aucune métrique acceptée, et partage l'unique objet état-apprenant local avec CDL et ZPD-Steer.
Nouveauté. Premier à fermer la boucle « bande de lutte dérivée du dialogue → commutation au décodage de la grammaire de coups socratiques », entièrement sur l'appareil. **Statut.** Contrôleur de commutation : implémenté & démontrable ; accord détecteur–enseignant : préliminaire ; gain d'apprentissage : proposé. **Validation.** Le contrôleur est démontrable aujourd'hui ; le gain d'apprentissage via essai randomisé pré/post + transfert différé (items isomorphes nouveaux), pré-enregistré, métrique = gain normalisé de Hake (g). Données de test restant sur l'appareil ; aucun chiffre de gain sans N pré-déclaré.

6 L'application en pratique

Notre assistant n'est pas un prototype de laboratoire : c'est une **application mobile réelle**, en ligne sur l'App Store et **téléchargée par ∼1 000 élèves**. La Fig. 7 en montre quatre écrans, toute l'inférence IA s'exécutant **100 % sur l'appareil**, sans qu'aucune donnée d'apprentissage ne quitte le téléphone.

7 Portabilité paneuropéenne et souveraineté

7.1 Un modèle de données canonique

L'Europe scolaire est fragmentée : chaque pays a ses systèmes (ENT) et son curriculum. Plutôt qu'une application par pays, nous projetons chaque ENT, via **un adaptateur unique**, sur un **modèle canonique commun** (`services/shared` : *Grade, Homework, Timetable, News, Attendance, Chat…*), au-dessus duquel toute la couche pédagogique est écrite *une seule fois* (Fig. 9). Le coût marginal d'un nouveau pays se réduit alors à deux briques : un adaptateur ENT et une spécialisation curriculaire (dataset + RAG dans la langue cible). Notre assistant embarque déjà **14 connecteurs**, dont 5 internationaux : ClasseViva (Italie, >2,5 M utilisateurs/jour), WebUntis (espace germanophone, >26 000 établissements), Moodle (>500 M utilisateurs), Canvas et Google Classroom ; Pronote couvre à lui seul le marché français (18 M utilisateurs, SecNumCloud 3.2). Au

cœur de cible européen, **37,5 M d'élèves du secondaire** de l'UE27, s'ajoute, via **Google Classroom**, le marché des **États-Unis** (∼49,5 M d'élèves K-12) [2], soit ∼**87 M d'élèves** (UE + USA) et **150 M d'utilisateurs** Google Classroom dans le monde. Sans infrastructure nouvelle, notre assistant se branche sur l'existant (Fig. 8).

7.2 Conformité européenne *par construction*

L'AI Act [16] classe l'éducation « à haut risque » (Annexe III.3), avec des obligations applicables dès le 2 août 2026. Le traitement local de l'inférence satisfait nativement gouvernance des données, minimisation et supervision humaine ; l'éducation restant « à haut risque » par défaut, la maïeutique de notre assistant (qui ne note ni n'oriente) relève de la dérogation de l'art. 6(3), tâche procédurale étroite, sans substitution au jugement enseignant, qu'il reste à confirmer. Côté RGPD (art. 25, *privacy by design*), **les données d'apprentissage (cours, conversations, notes) ne quittent pas l'appareil** ; seule une progression non nominative (XP, niveau) est synchronisée pour le classement, sans contenu scolaire personnel exposé au *US CLOUD Act*. L'enjeu est aussi stratégique : le rapport Draghi (2024) relève que seules **4 des 50** premières entreprises technologiques mondiales sont européennes, et l'initiative *EuroStack* (∼300 Md€) prône un principe « *open source first* ». Là où Gaia-X illustre les limites d'une souveraineté « par le haut » (bureaucratie, capture par les hyperscalers), notre assistant incarne une souveraineté **« par le bas »**, distribuée sur des millions d'appareils *déjà* possédés et bâtie sur des briques ouvertes (Gemma, `llama.cpp`/GGUF, EmbeddingGemma, dataset libre) : pas de datacenter à bâtir, la souveraineté tient dans la poche de l'élève (Tab. 5).

8 Résultats

8.1 Vitesse et frugalité (H3, H6)

La génération locale est opérationnelle sur tous les appareils testés (Tab. 3), de 30 tok/s sur iPhone 14 à 7 tok/s sur les tablettes Lenovo distribuées par la Région Sud, lent mais utilisable pour un tutorat où l'élève lit chaque réponse. Surtout, l'inférence est **frugale** : ∼3 W, soit ∼0,005 Wh/requête (≈3 W × 6 s) contre ∼0,24–0,30 Wh pour une requête cloud [17], et jusqu'à ∼33 Wh pour un modèle de *raisonnement* (Fig. 12), soit ∼50 à ∼6 600 fois moins (énergie d'usage ; fabrication et recharge des appareils, déjà possédés, non comptées). Le contexte rend l'écart décisif : les datacenters ont consommé ∼415 TWh en 2024 (∼1,5 % de l'électricité mondiale), ∼945 TWh projetés en 2030 (AIE), et un email de 100 mots rédigé par ChatGPT mobiliserait ∼519 mL d'eau (UC Riverside, 2025). Consommation **mesurée au wattmètre** en génération continue (Fig. 10).

8.2 Qualité pédagogique (H1)

Sur 40 questions notées à l'aveugle par 2 correcteurs (Fig. 13), ChatGPT excelle, comme attendu, en pertinence (4,5) et rigueur (4,2) ; mais conçu pour livrer la réponse, il échoue en posture pédagogique (**1,6**) et en adaptation au programme (2,9). Surtout, le modèle de base brut (Llama 3.2) obtient lui aussi une posture faible (**1,2**) que notre *fine-tuning* porte à **4,6** : c'est la spécialisation, non la puissance du modèle, qui installe la posture. Sur les **quatre critères pédagogiques**, notre assistant obtient **4,2/5 contre 3,3** pour ChatGPT (le critère structurel *hors-ligne*, 0 vs 5, déjà couvert par **H3**, est traité à part). Ces écarts restent une *tendance robuste* (κ/IC à calculer), non une preuve définitive.

8.3 Confiance et adoption (H2)

Au questionnaire ($N = 47$), **89 %** recommanderaient notre assistant, **74 %** le préfèrent à ChatGPT (« ça oblige à réfléchir »), **78 %** rapportent une meilleure rétention. Sur la vie privée,

2. Eurostat 2024 (UE, CITE 2+3) ; NCES 2024 (K-12 USA) ; Google for Education, 2025.

l'écart est net : **87 %** font confiance à notre assistant (4,4/5) contre **19 %** à ChatGPT (2,3/5), le traitement local est compris comme un avantage concret. Six mois après le lancement, l'application a été **téléchargée par ∼1 000 élèves** (dont ∼250 utilisateurs actifs) (Fig. 11). Le Tab. 4 synthétise la validation des hypothèses.

8.4 CDL, un proxy embarqué de la « dette cognitive » (H1)

Le *Cognitive Debt Ledger* calcule, à partir de la seule trace de conversation déjà locale (WatermelonDB) et des comptes de tokens du moteur, un vecteur comportemental par tour, *sans capteur ni donnée quittant l'appareil*. Signal central : le **Student Reasoning Ratio** au tour τ,

$$\mathrm{SRR}_\tau = \frac{n^{\text{rais.}}_{\text{élève},\tau}}{n^{\text{rais.}}_{\text{élève},\tau} + n^{\text{rais.}}_{\text{tuteur},\tau}}, \tag{9}$$

où $n^{\text{rais.}}$ compte les tokens de raisonnement, complété du taux de demande-de-réponse $\rho_{\text{dem.}}$, du taux de tentative-avant-indice $\rho_{\text{tent.}}$ et de la latence avant première tentative. Agrégés sur la session, ces signaux forment un *indice de dette cognitive*

$$\mathrm{CDI} = w_1\big(1 - \overline{\mathrm{SRR}}\big) + w_2\,\rho_{\text{dem.}} + w_3\big(1 - \rho_{\text{tent.}}\big), \qquad \textstyle\sum_k w_k = 1. \tag{10}$$

La séparation des tokens de raisonnement *réutilise la grammaire de coups GBNF de PEDAL* (éq. (3)) qui segmente déjà un tour du tuteur en {diagnostic | indice | question} ; les tours de l'élève sont classés par une tête logistique légère sur les plongements EmbeddingGemma déjà résidents, sans téléchargement de modèle. Une baisse soutenue du SRR fait *escalader* le tuteur vers le questionnement, fermant la boucle que les *benchmarks* se contentent de mesurer. Nous ne prétendons pas que le CDL *mesure* la dette cognitive : c'est un *proxy comportemental*, distinct du construit neuronal mesuré par EEG [9].

Nouveauté. Premier proxy de *décharge cognitive* entièrement embarqué (aucune égression), fondé sur la grammaire de coups pour isoler les tokens de raisonnement du tuteur. **Statut.** Proposé (conçu, non validé) ; fiabilité d'étiquetage : mesurable ; corrélation SRR↔offloading : préliminaire. **Validation.** Étude de validité de construit pré-enregistrée sur la cohorte $N = 47$: corrélation (Pearson & Spearman) entre SRR et items d'offloading auto-reportés, *conditionnée* à un κ inter-annotateurs acceptable. $N = 47$ étant petit, un résultat nul est une issue probable et honnête.

9 Impact, limites et éthique

Impact. Notre assistant agit sur trois fractures européennes. *Équité* : sans abonnement ni connexion, sur une tablette à 150 €, il atteint les élèves que le cloud abandonne (zones peu connectées, milieux modestes). *Écologie* : à l'échelle des 37,5 M lycéens, ∼10 interactions/jour de classe représentent $\approx 6{,}8 \times 10^{10}$ requêtes/an. En basculant ces requêtes du cloud (∼0,3 Wh ; ∼0,3 mL d'eau, estimation prudente par requête courte) vers l'embarqué, on évite chaque année, en ordre de grandeur :

$$\underbrace{6{,}8 \times 10^{10} \times 0{,}3\,\mathrm{Wh}}_{\approx 20\ \mathrm{GWh}} \quad \text{et} \quad \underbrace{6{,}8 \times 10^{10} \times 0{,}3\,\mathrm{mL}}_{\approx 20\,000\ \mathrm{m^3\ d'eau}},$$

soit la consommation électrique de ∼5 000 foyers européens et le volume de ∼8 piscines olympiques. *Souveraineté* : une IA qui appartient à l'élève et reste en Europe. Ces axes recoupent les ODD 4, 10, 12 et 13 de l'ONU (Fig. 14). En tournant sur des appareils anciens, notre assistant prolonge en outre leur durée de vie, une sobriété matérielle qui s'ajoute à la sobriété énergétique.

Éthique. Construire une IA pour des mineurs impose une vigilance accrue : les biais des modèles (atténués par la validation enseignante et l'ancrage RAG sur des manuels officiels), le risque d'une *nouvelle* dépendance (contré *par conception* par l'absence de réponse directe et la friction socratique), et la transparence (dataset publié sous licence ouverte, demande de brevet déposée sur l'architecture auprès de l'INPI). Aucune finalité commerciale n'a guidé la conception scientifique.

Limites et biais. Échantillon réduit (47 élèves, un établissement ; validité externe limitée, étude multi-établissements $n > 200$ prévue) ; biais de sélection (volontaires) ; un correcteur est notre encadrant (évaluation à jury élargi prévue) ; modèles >4B incompatibles avec <3 Go de RAM ; spécialisation curriculaire complète pour la France seulement, les autres pays relevant de la feuille de route. La « rétention perçue » déclarée n'est pas une mesure objective d'apprentissage et appelle un protocole pré/post-test contrôlé.

Honnêteté. L'étude MIT que nous citons a un effectif modeste et une méthodologie EEG discutée [10] : nous la mobilisons comme *motivation*, non comme preuve. Notre dataset publié compte 1 000 paires *validées* (938 « gold »), retenues sur ∼10 000 candidats générés. Le RAG réduit les hallucinations sans les supprimer ; la posture socratique, qui questionne plutôt qu'elle n'affirme, en limite mécaniquement la portée.

10 Conclusion

Notre assistant démontre, mesures à l'appui, qu'une IA éducative peut *amplifier* le raisonnement au lieu de s'y substituer, tout en restant privée (87 % de confiance vs 19 %), frugale (∼3 W) et souveraine, et portable à l'échelle d'un continent via un modèle de données canonique. Perspectives : achever l'orchestrateur local⇄cloud, le tutorat multimodal (vision, voix), l'extension du dataset/RAG à d'autres curricula et langues européens, et un benchmark ouvert précision/-énergie de l'IA éducative embarquée. *La meilleure IA n'est pas celle qui donne la réponse, c'est celle qui aide à la trouver.*

Écosystème partenaire. **Polytech Nice Sophia** nous a guidés sur l'architecture hybride et l'optimisation embarquée ; les **professeurs du Lycée Henri Matisse** ont conçu et validé le corpus ; **AxePI** a protégé l'innovation (brevet INPI) et hébergé le serveur ; **Lucarne Pro** a publié l'app sur l'App Store ; **Fondation CGénial** et **Sciences à l'École** ont encadré et qualifié le projet pour EUCYS.

> *« À l'heure où tant d'élèves se contentent de demander la réponse à une IA, ceux-ci ont fait l'inverse : ils ont voulu remettre de l'humain dans le projet, le construire avec nous, les professeurs, et chercher toujours plus loin, au lieu de rester dans cette « génération IA » qui ne réfléchit plus. »*
> , Carolle Chabanier, professeure de mathématiques, Lycée Henri Matisse

Reconnaissance, Prix Orange Numérique

Notre assistant a reçu le **Prix Orange Numérique** (Concours Science Factor 2026). Ce soutien ouvre la voie à un passage du prototype vers un déploiement à plus grande échelle, dans une logique de numérique éducatif responsable, inclusif et souverain.

Remerciements. Merci à M. P. Aboussouan, à nos partenaires, aux lycéens testeurs et à nos familles. Notre assistant est **lauréat national du Concours C'Génial 2026** (qualification EUCYS) et **double lauréat du Concours Science Factor 2026**.

Supplément (document séparé) : protocole d'évaluation détaillé (grille, contrôle des biais, traitement statistique) et construction du corpus Lucid-Socratic-EDU.