

Figure 1. Résumé graphique : du problème (la « dette cognitive ») à un tuteur socratique embarqué et souverain, et mesuré.

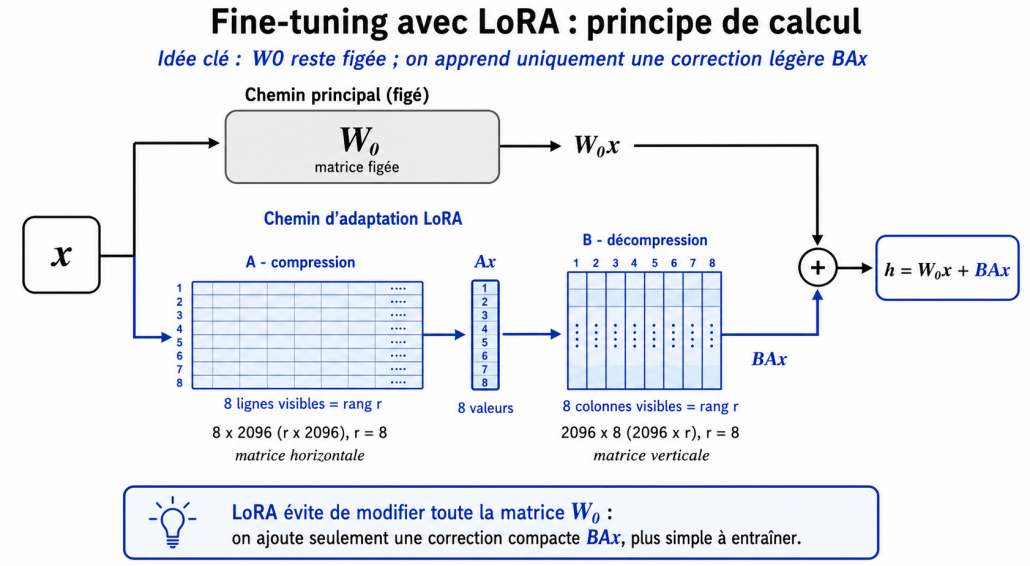


Figure 2. Principe du *fine-tuning* LoRA/QLoRA (éq. 2) : le modèle de base W_0 reste *figé* ; on n'apprend qu'une correction de rang faible $\Delta W\, x = BA\, x$ (schéma de principe à rang $r = 8$; configuration retenue $r = 16$), soit ∼0,1 % des paramètres, « un petit carnet de notes » qui corrige le seul comportement pédagogique.

Exemple de dialogue socratique (mathématiques)

Élève — Je dois résoudre $x^2 - 5x + 6 = 0$ mais je ne sais pas comment.
Lucid — De quel degré est cette équation ? Quelle méthode connais-tu ?
Élève — Degré 2…le discriminant ?
Lucid — Que vaut $\Delta = b^2 - 4ac$ avec a=1, b=−5, c=6 ?
Élève — $25 - 24 = 1$.
Lucid — $\Delta > 0$: combien de solutions cela annonce-t-il, et avec quelle formule ?

Figure 3. Le tuteur ne livre jamais la réponse (x=2, 3) : il y conduit par paliers.

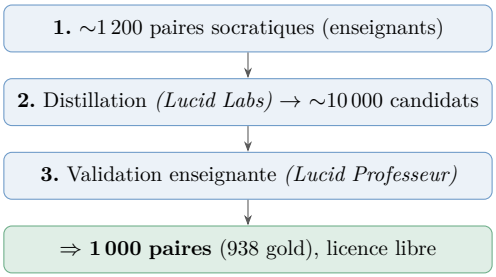


Figure 4. Entonnoir du dataset : de ∼10 000 candidats générés à 1 000 paires validées et ouvertes (938 « gold », 18 matières, 7 niveaux).

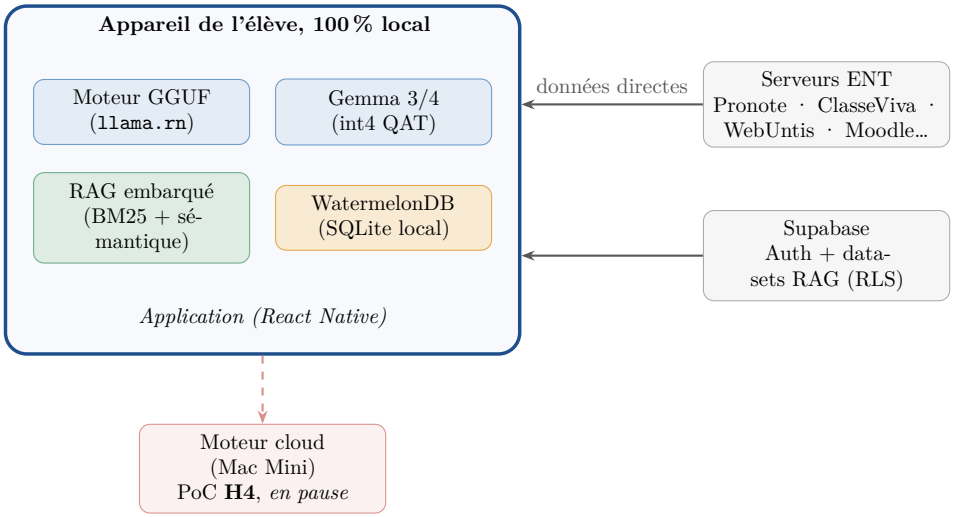


Figure 5. Architecture *Local-First*. Toute l'inférence s'exécute sur l'appareil ; les données ENT y sont rapatriées directement, sans intermédiaire. Le moteur cloud (**H4**) reste une preuve de concept en pause.

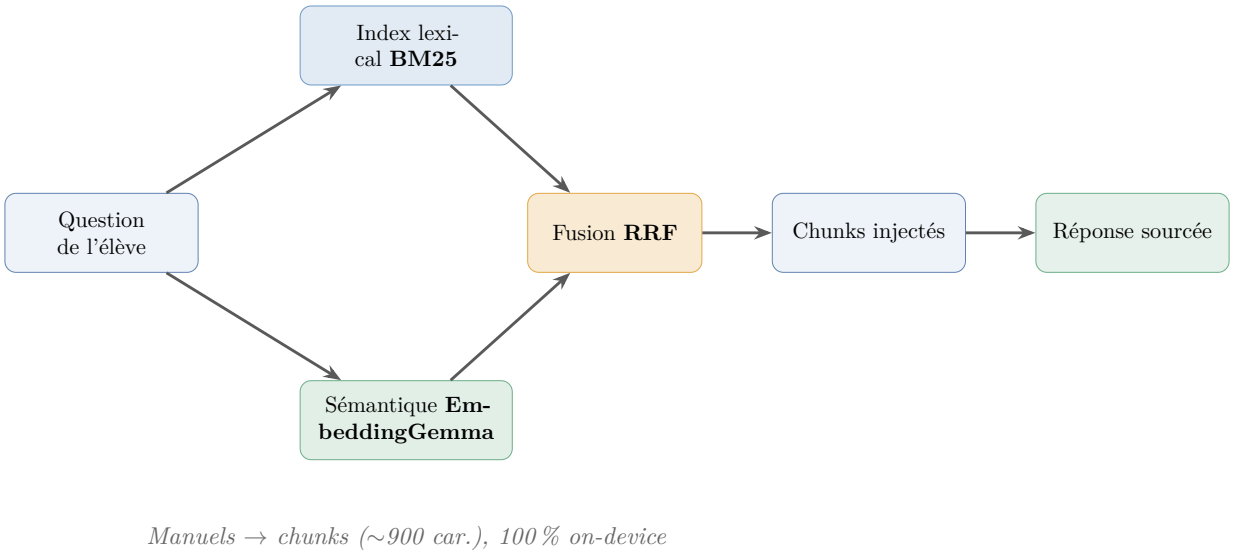
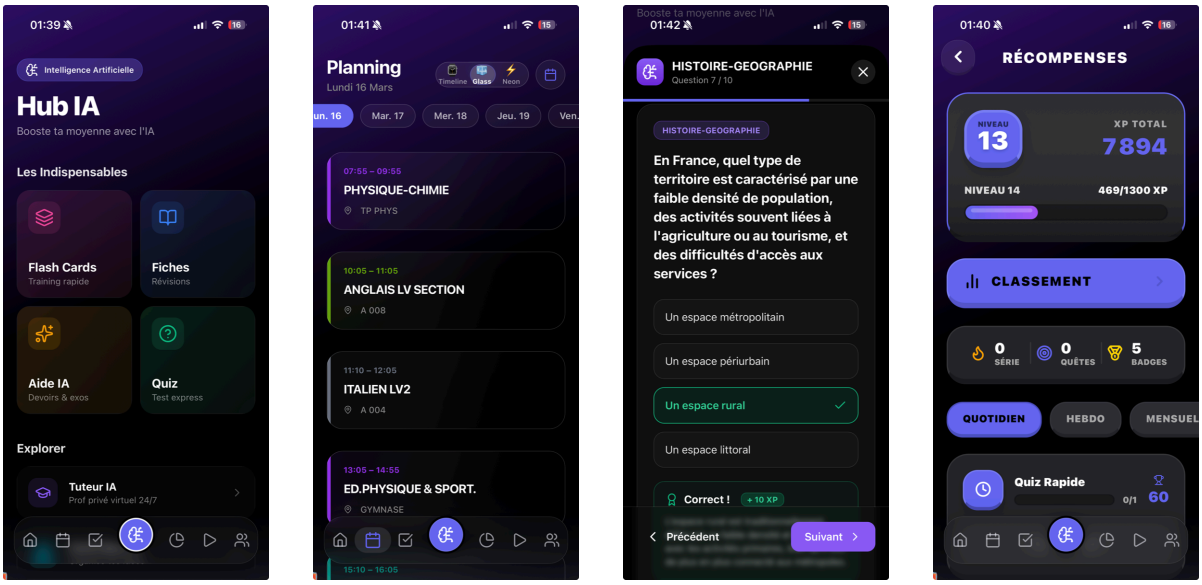


Figure 6. RAG hybride embarqué : recherches lexicale (BM25) et sémantique (EmbeddingGemma) fusionnées par *Reciprocal Rank Fusion*, intégralement sur l'appareil.



Hub IA **Emploi du temps** **Quiz du cours** **Gamification**

Figure 7. notre assistant en pratique (captures réelles). **1.** Le *Hub IA* : flashcards, fiches, quiz, aide et *tuteur socratique*. **2.** L'*emploi du temps* rapatrié directement de l'ENT (ici Pronote). **3.** Un *quiz* généré à partir du cours réel de l'élève, avec correction immédiate et XP. **4.** La *gamification* de l'effort (niveaux, classement, badges), pour récompenser le travail, pas la réponse.

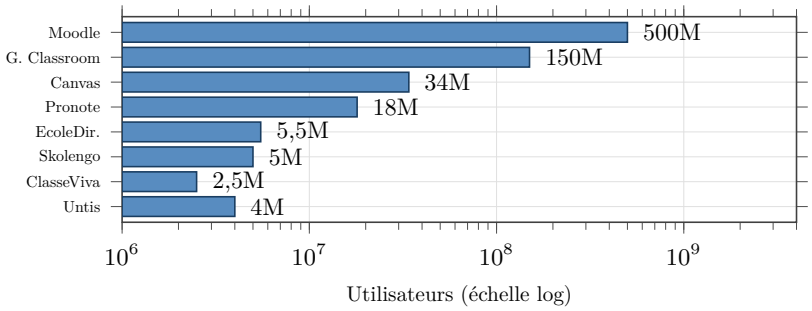


Figure 8. Portée des systèmes scolaires accessibles via les connecteurs de notre assistant (sources : éditeurs, 2024–2026).

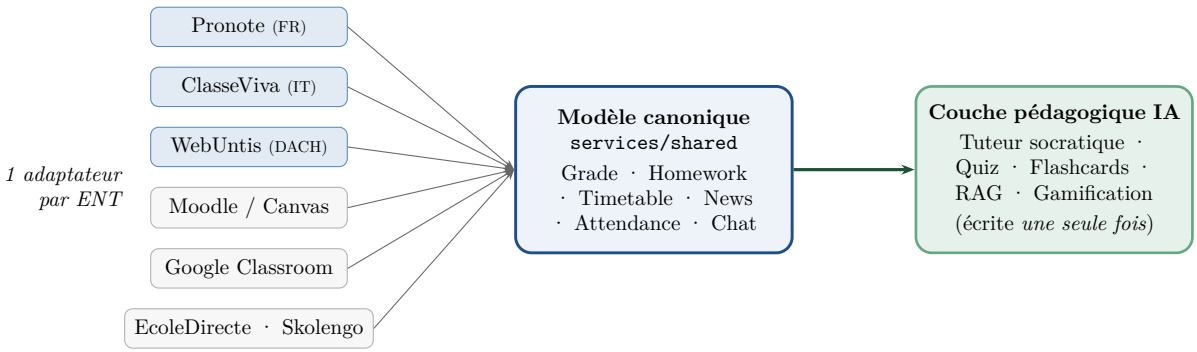


Figure 9. Modèle de données canonique : chaque ENT est projeté par un adaptateur unique sur un modèle commun, au-dessus duquel l'intelligence pédagogique est mutualisée. Ajouter un pays ≈ ajouter un adaptateur (→**H5**).

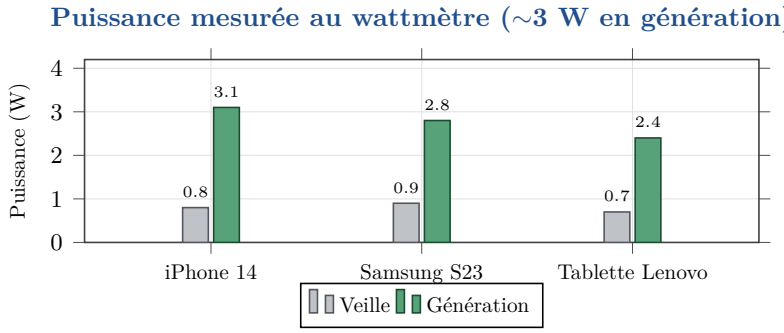


Figure 10. Puissance électrique mesurée au wattmètre, en veille et pendant la génération (médiane sur 30 requêtes). La surconsommation imputable à l'IA reste de ∼2–3 W, l'ordre de grandeur d'une ampoule LED.

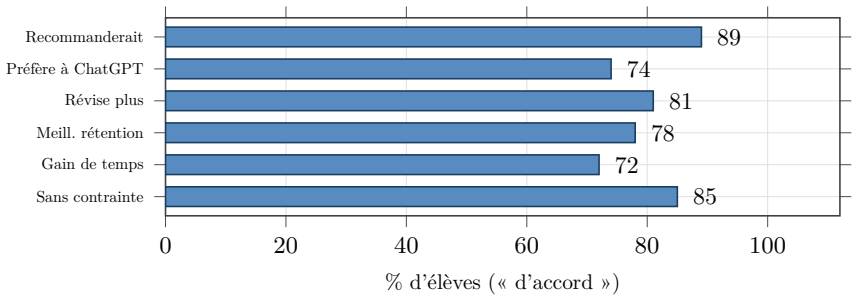


Figure 11. Questionnaire ($N = 47$ lycéens). 87 % font confiance à notre assistant pour leurs données, contre 19 % à ChatGPT.

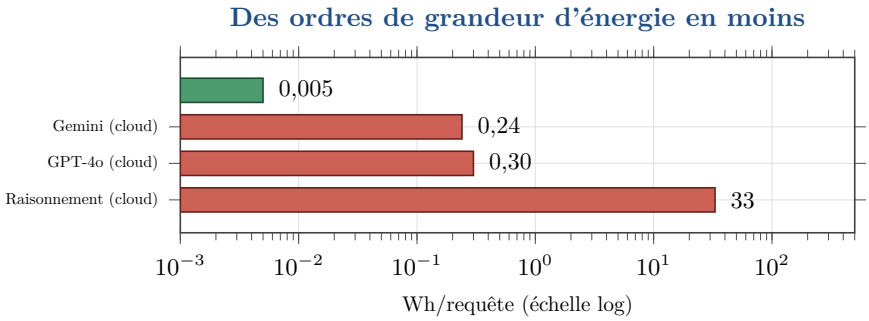


Figure 12. Énergie par requête : l'embarqué consomme ∼50 à ∼6 600× moins qu'une IA cloud.

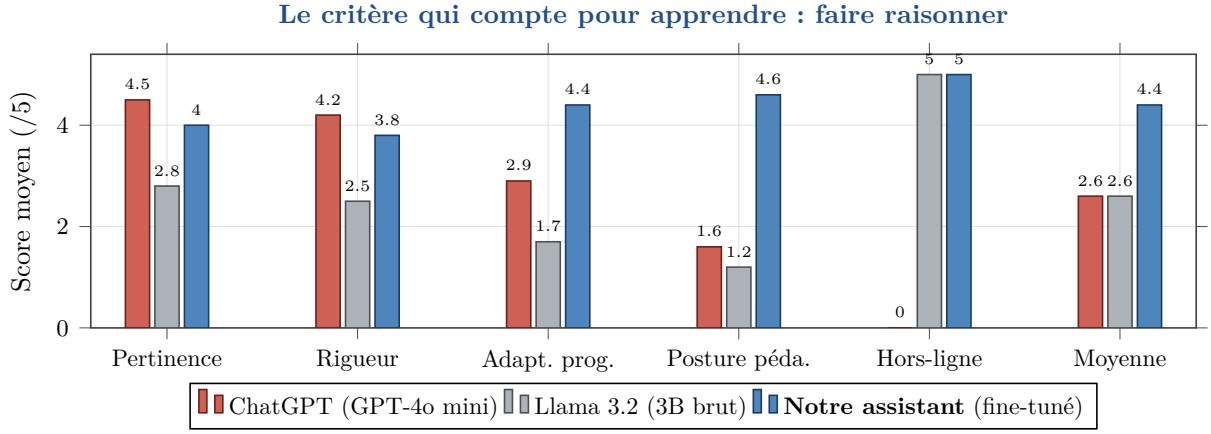


Figure 13. Évaluation comparative à l'aveugle (5 critères, 40 questions, 2 correcteurs). Le *fine-tuning* installe la posture socratique (4,6) là où ChatGPT (1,6) et le modèle brut Llama 3.2 (1,2) livrent la réponse ; sur les 4 critères pédagogiques, 4,2 vs 3,3 (le hors-ligne, structurel, est exclu).

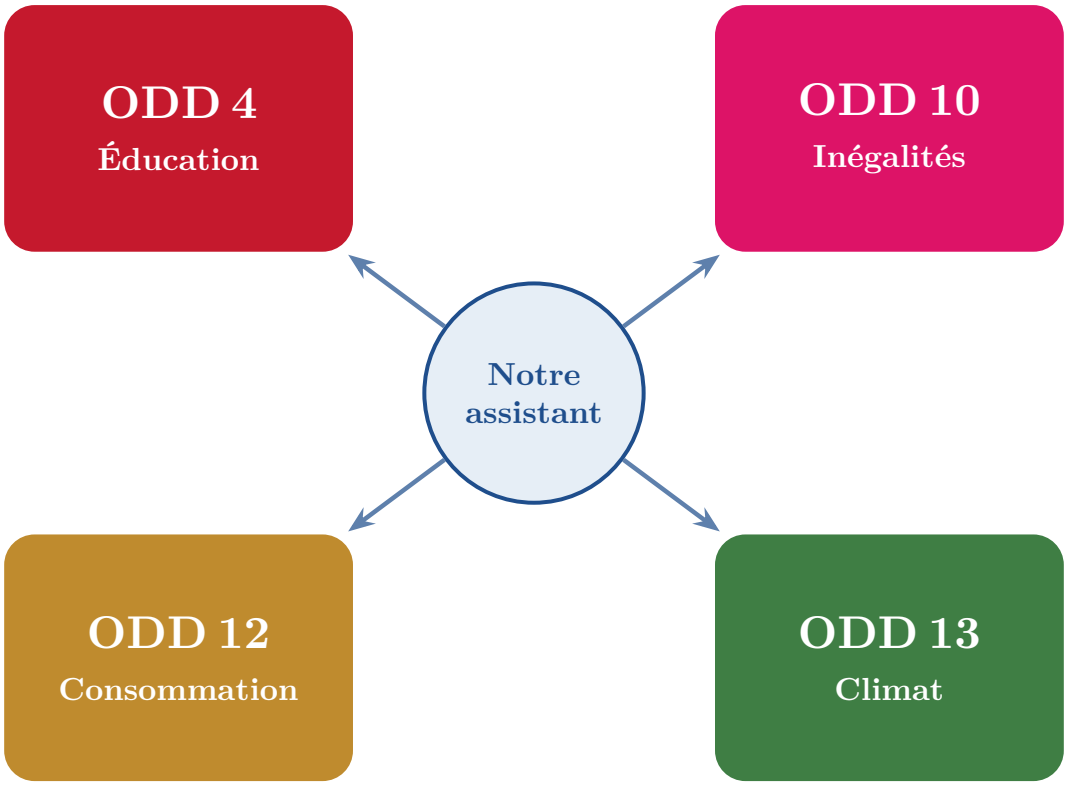


Figure 14. Alignement de notre assistant avec les Objectifs de développement durable de l'ONU.

Table 1. Positionnement de notre assistant face aux outils existants (à notre connaissance). ✓ : oui ; **p** : partiel ; × : non.

Propriété	**notre assistant**	ChatGPT	Photomath	Khanmigo	Modèle local brut
Posture socratique (ne livre pas la réponse)	✓	×	×	**p**	×
Traitement 100 % local (privé)	✓	×	×	×	✓
Fonctionne hors-ligne	✓	×	×	×	✓
Ancré aux programmes (RAG + fine-tuning)	✓	**p**	**p**	✓	×
Souverain (UE, RGPD *by-design*)	✓	×	×	×	✓
Frugal (∼3 W, sans datacenter)	✓	×	×	×	✓

Table 2. Les six contributions originales de notre assistant (⋆ : contribution phare). Statut : i implémenté & mesuré, p prototype/préliminaire, o proposé (protocole pré-enregistré). Chaque revendication de nouveauté a été vérifiée contre la littérature comme *composition inédite* de primitives connues.

Contribution	**Principe & nouveauté (« premier à... »)**	**Statut**
⋆ **PAQ**	*Posture-Aware Quantization* : l'allocation des bits int4 pilotée par une marge comportementale *socratique-vs-réponse* (signal W-Gap), préservant la posture que la quantification standard érode, à taille et latence identiques.	p
PEDAL	Grammaire GBNF à deux couches imposant la structure des coups socratiques *et* rendant la réponse littérale grammaticalement inatteignable au décodage, sur LLM int4 embarqué.	p
ZED / W-Gap	La réponse d'or (locale) comme contrainte négative au décodage ; garde de fuite *et* métrique (W-Gap) entièrement *on-device*.	i+p
ZPD-Steer	L'habileté θ (estimation type Elo, apparentée à Rasch) estimée sur l'appareil pilote le reclassement RAG *et* la génération, sans exfiltration.	p
CDL / SRR	*Student Reasoning Ratio* : proxy embarqué de la « dette cognitive », fondé sur la grammaire de coups.	o
PS2	Bande de lutte productive dérivée du dialogue → commutation de la grammaire socratique au décodage.	i+p

Table 3. Débit de génération (médiane sur 30 requêtes ; variable selon le modèle embarqué, E2B/4B, et la longueur de réponse).

Configuration	**tok/s**	**Données**
Embarqué, iPhone 14 (A15)	30	on-device
Embarqué, Samsung S23	24	on-device
Embarqué, Redmi Note 12	11	on-device
Embarqué, tablette Lenovo	7	on-device
ChatGPT (GPT-4o mini, API)	110	serveurs US

Table 4. Validation des hypothèses.

Hypothèse	Résultat	Statut
H1 Socratique	Posture 4,6/5 par *fine-tuning* (vs 1,6 ChatGPT, 1,2 Llama brut ; n=40, κ/IC à calculer) ; 78 % meilleure rétention *perçue*.	Soutenue
H2 Confiance	87 % de confiance vs 19 % ; traitement local perçu comme rassurant.	Validée
H3 Hors-ligne	Génération locale sur tous les appareils (30 à 7 tok/s) ; 85 % « sans contrainte ».	Validée
H4 Orchestration	PoC validé (<800 ms) mais mis en pause en production (infrastructure).	En pause
H5 Portabilité	14 connecteurs (dont 5 internationaux) sur un modèle canonique unique.	Validée (archi.)
H6 Frugalité	∼3 W ; ∼0,005 Wh/requête vs ∼0,24–0,30 Wh cloud (∼50–6 600× moins).	Validée

Table 5. Conformité *par construction* aux cadres européens.

Cadre	Exigence	Réponse ***by-design*** de notre assistant
AI Act (éducation, haut risque)	Gouvernance, supervision humaine, transparence	Traitement local ; ne note ni n'oriente ; tuteur supervisable
RGPD art. 25	*Privacy by design*, minimisation	Données d'apprentissage 100 % locales ; seule la progression non nominative (XP) synchronisée
RGPD art. 8 (mineurs)	Consentement variable, 27 régimes	Aucun contenu scolaire personnel transféré ; base légale du compte à documenter
US CLOUD Act vs RGPD	Risque d'accès extraterritorial	Inférence et contenu scolaire locaux ⇒ contenu sensible non exposé
Green Deal	Sobriété numérique	∼3 W, pas de datacenter

Références

[1] J. Sweller (1988). Cognitive Load During Problem Solving. *Cognitive Science* 12(2).

[2] H. Ebbinghaus (1885). *Über das Gedächtnis*.

[3] N. Cepeda *et al.* (2006). Distributed Practice : A Review and Quantitative Synthesis. *Psychological Bulletin* 132(3).

[4] H. Roediger, J. Karpicke (2006). Test-Enhanced Learning. *Psychological Science* 17(3).

[5] R. Ryan, E. Deci (2000). Self-Determination Theory. *American Psychologist* 55(1).

[6] E. Kasneci *et al.* (2023). ChatGPT for Good ? *Learning and Individual Differences* 103.

[7] G. Kestin *et al.* (2024). AI Tutoring Outperforms In-Class Active Learning. Harvard (N=194).

[8] L. Deslauriers *et al.* (2019). Measuring actual learning vs feeling of learning. *PNAS* 116(39).

[9] N. Kosmyna *et al.* (2025). *Your Brain on ChatGPT*. MIT Media Lab, arXiv :2506.08872.

[10] M. Stankovic *et al.* (2026). Comment on « Your Brain on ChatGPT ». arXiv :2601.00856.

[11] Gemma Team, Google DeepMind (2024). *Gemma 2*. arXiv :2408.00118.

[12] Gemma Team, Google DeepMind (2025–2026). *Gemma 3 & 4 QAT : Quantization-Aware Training for State-of-the-Art On-Device AI*. Google Developers Blog.

[13] T. Dettmers *et al.* (2023). *QLoRA*. NeurIPS 36.

[14] E. Hu *et al.* (2021). *LoRA*. arXiv :2106.09685.

[15] P. Lewis *et al.* (2020). *Retrieval-Augmented Generation*. NeurIPS 33.

[16] Règlement (UE) 2024/1689 (AI Act). EUR-Lex.

[17] S. Luccioni, Y. Jernite, E. Strubell (2024). Power Hungry Processing. *ACM FAccT*.

[18] K. Yang, D. Klein (2021). FUDGE : Controlled Text Generation With Future Discriminators. *NAACL*.

[19] X. Lu *et al.* (2021). NeuroLogic Decoding : (Un)supervised Neural Text Generation with Predicate Logic Constraints. *NAACL*.

[20] G. Rasch (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Pædagogiske Institut.

[21] L. S. Vygotsky (1978). *Mind in Society : The Development of Higher Psychological Processes*. Harvard University Press.

[22] R. A. Bjork (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (eds.), *Metacognition*, MIT Press.

[23] J. E. Beck, Y. Gong (2013). Wheel-Spinning : Students Who Fail to Master a Skill. *AIED*.

[24] J. Lin *et al.* (2024). AWQ : Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration. *MLSys*.