

Figure 1. Visual abstract: from the problem (the « cognitive debt ») to an on-device, sovereign Socratic tutor, and measured.

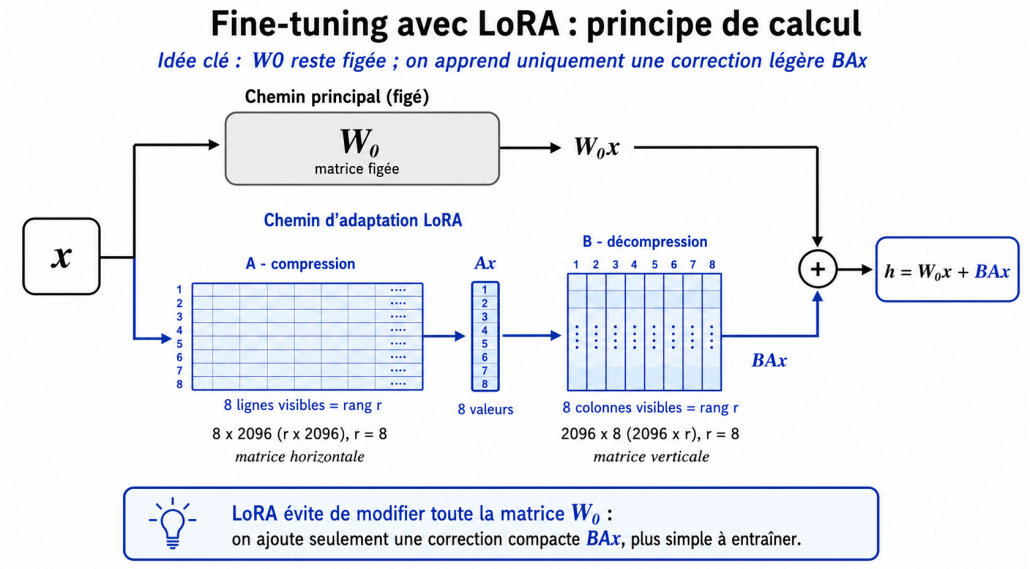


Figure 2. Principle of LoRA/QLoRA *fine-tuning* (eq. 2): the base model W_0 remains *frozen*; only a low-rank correction $\Delta W\, x = BA\, x$ is learned (schematic at rank $r = 8$; retained configuration $r = 16$), i.e. ∼0.1% of the parameters, "a small notebook" that corrects only the pedagogical behavior.

Example of Socratic dialogue (mathematics)

Student — I have to solve $x^2 - 5x + 6 = 0$ but I don't know how.
Lucid — What degree is this equation? What method do you know?
Student — Degree 2...the discriminant?
Lucid — What is $\Delta = b^2 - 4ac$ with a=1, b=−5, c=6?
Student — $25 - 24 = 1$.
Lucid — $\Delta > 0$: how many solutions does that announce, and with which formula?

Figure 3. The tutor never delivers the answer (x=2, 3): it leads to it in stages.

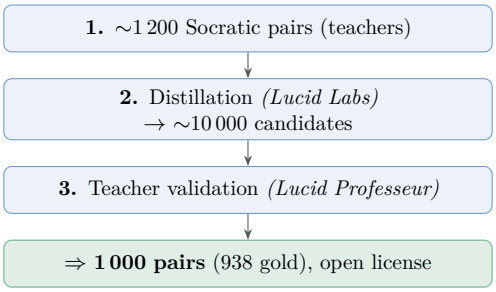


Figure 4. Dataset funnel: from ∼10 000 generated candidates to 1 000 validated and open pairs (938 "gold", 18 subjects, 7 levels).

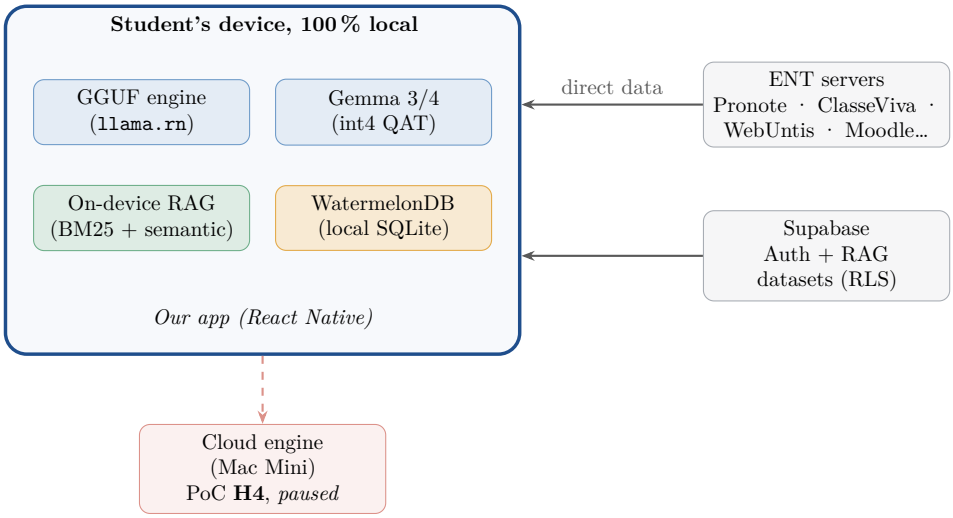


Figure 5. *Local-First* architecture. All inference runs on the device; ENT data is brought back to it directly, without an intermediary. The cloud engine (**H4**) remains a paused proof of concept.

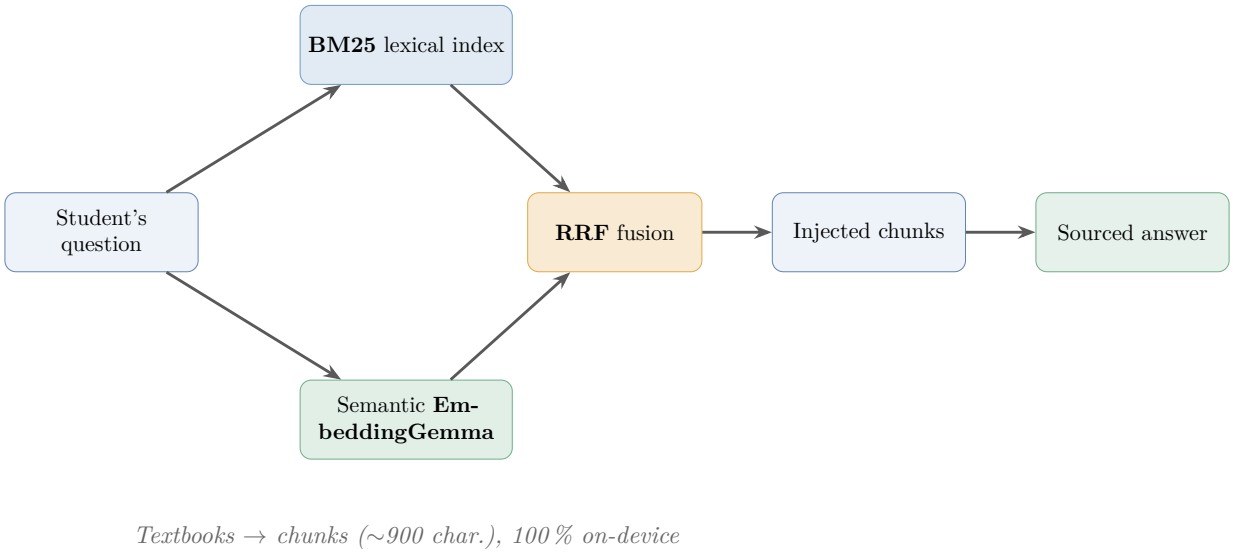
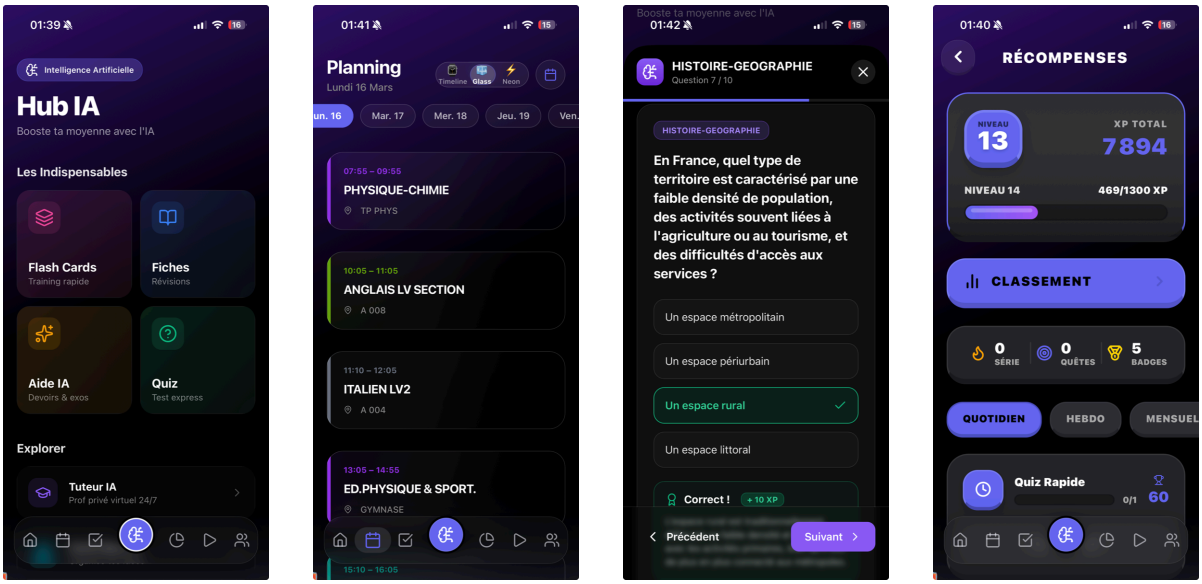


Figure 6. On-device hybrid RAG: lexical (BM25) and semantic (EmbeddingGemma) searches fused by *Reciprocal Rank Fusion*, entirely on the device.



AI Hub **Timetable** **Course quiz** **Gamification**

Figure 7. our app in practice (real screenshots). **1.** The *AI Hub*: flashcards, revision sheets, quizzes, help and *Socratic tutor*. **2.** The *timetable* fetched directly from the school digital workspace (ENT) (here Pronote). **3.** A *quiz* generated from the student's real course material, with immediate grading and XP. **4.** The *gamification* of effort (levels, leaderboard, badges), to reward the work, not the answer.

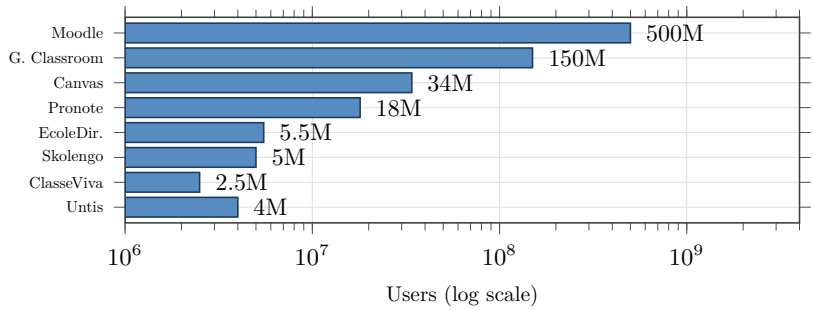


Figure 8. Reach of the school systems accessible through our app's connectors (sources: vendors, 2024–2026).

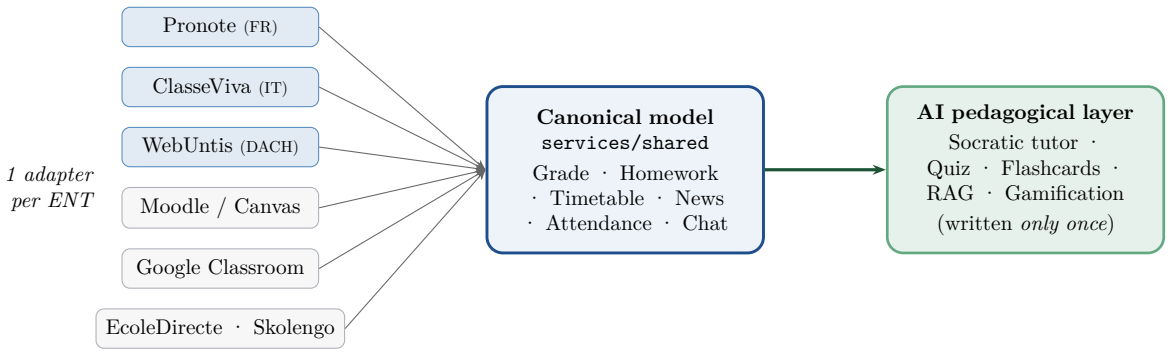


Figure 9. Canonical data model: each ENT is projected by a single adapter onto a shared model, on top of which pedagogical intelligence is shared. Adding a country ≈ adding an adapter (→**H5**).

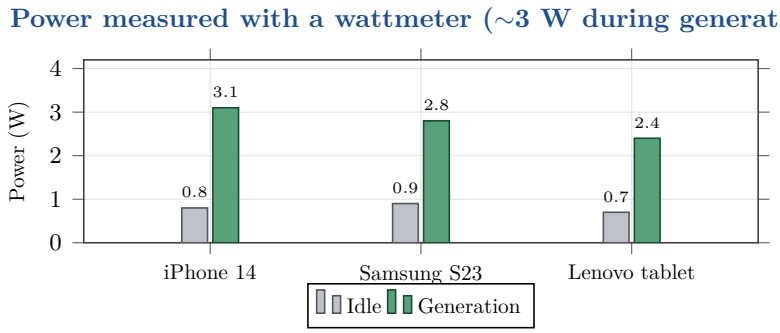


Figure 10. Electrical power measured with a wattmeter, when idle and during generation (median over 30 requests). The over-consumption attributable to AI remains ∼2–3 W, the order of magnitude of an LED bulb.

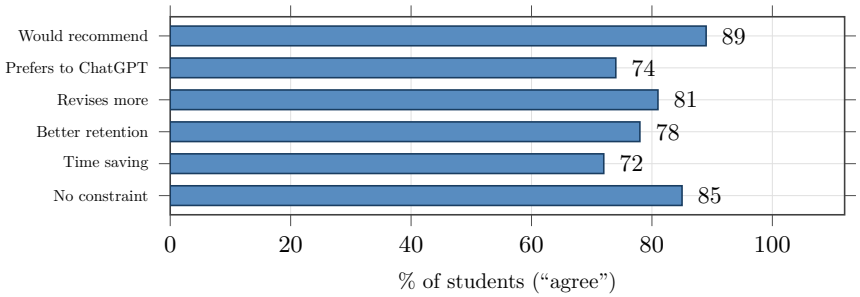


Figure 11. Survey ($N = 47$ secondary-school students). 87 % trust our app with their data, against 19 % for ChatGPT.

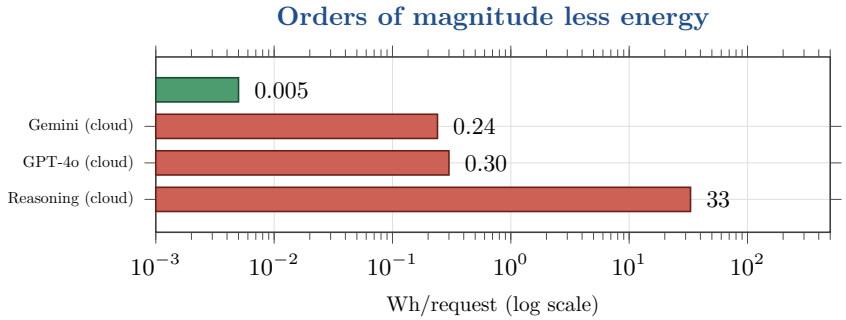


Figure 12. Energy per request: on-device consumes ∼50 to ∼6 600× less than a cloud AI.

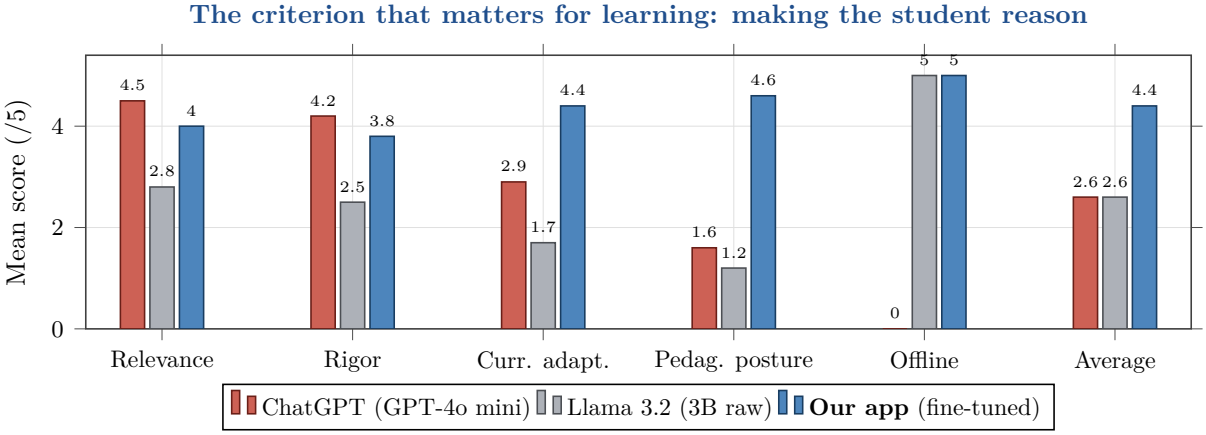


Figure 13. Blind comparative evaluation (5 criteria, 40 questions, 2 graders). Fine-tuning installs the Socratic posture (4.6) where ChatGPT (1.6) and the raw Llama 3.2 model (1.2) deliver the answer; on the 4 pedagogical criteria, 4.2 vs 3.3 (the offline criterion, structural, is excluded).

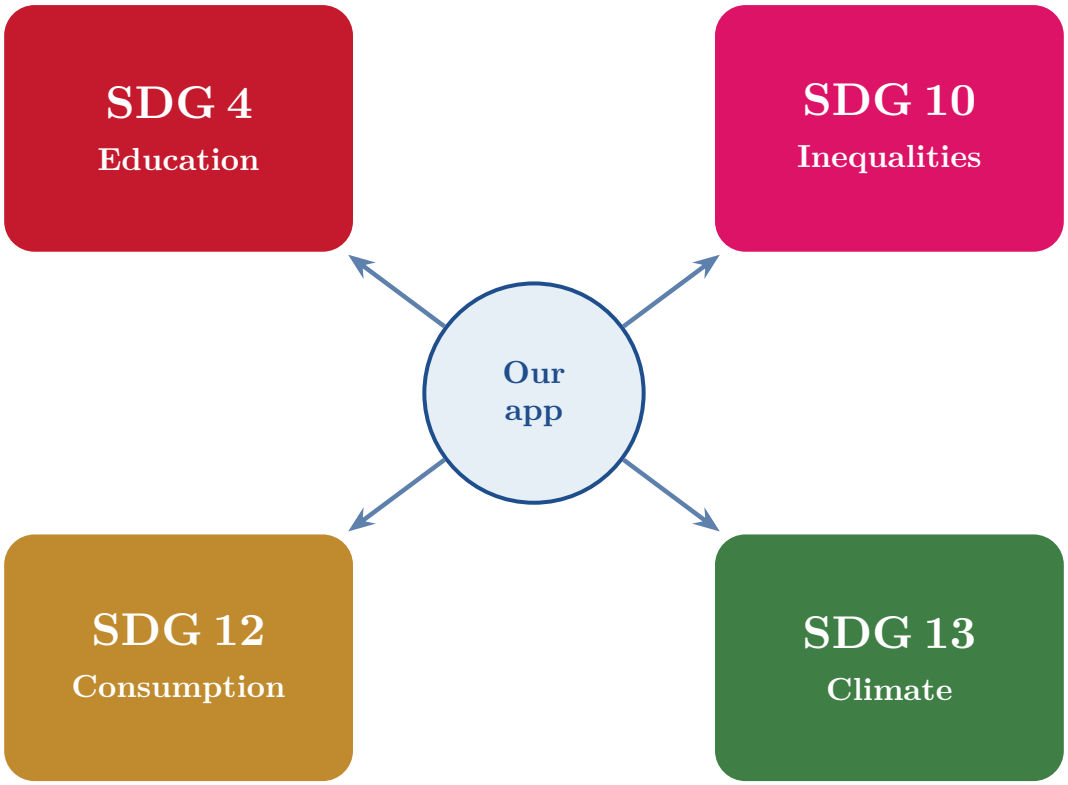


Figure 14. Alignment of our app with the UN Sustainable Development Goals.

Table 1. Positioning of our app against existing tools (to our knowledge). ✓ : yes ; **p** : partial ; × : no.

Property	**our app**	ChatGPT	Photomath	Khanmigo	Raw local model
Socratic posture (does not hand over the answer)	✓	×	×	**p**	×
100 % local processing (private)	✓	×	×	×	✓
Works offline	✓	×	×	×	✓
Grounded in curricula (RAG + fine-tuning)	✓	**p**	**p**	✓	×
Sovereign (EU, GDPR *by-design*)	✓	×	×	×	✓
Frugal (∼3 W, without datacenter)	✓	×	×	×	✓

Table 2. The six original contributions of our app (⋆ : flagship contribution). Status: i implemented & measured, p prototype/preliminary, o proposed (pre-registered protocol). Each novelty claim has been verified against the literature as a *novel composition* of known primitives.

Contribution	**Principle & novelty ("first to...")**	**Status**
⋆**PAQ**	*Posture-Aware Quantization*: int4 bit allocation driven by a *Socratic-vs-answer* behavioral margin (W-Gap signal), preserving the posture that standard quantization erodes, at identical size and latency.	p
PEDAL	Two-layer GBNF grammar enforcing the structure of Socratic moves *and* rendering the literal answer grammatically unreachable at decoding, on an on-device int4 LLM.	p
ZED / W-Gap	The (local) gold answer as a negative constraint at decoding; leakage guard *and* metric (W-Gap) entirely *on-device*.	i+p
ZPD-Steer	The ability θ (Elo-type estimate, related to Rasch) estimated on-device drives both RAG reranking *and* generation, without exfiltration.	p
CDL / SRR	*Student Reasoning Ratio*: on-device proxy of "cognitive debt", grounded in the move grammar.	o
PS2	Productive struggle band derived from the dialogue → switching of the Socratic grammar at decoding.	i+p

Table 3. Generation throughput (median over 30 requests; varies with the on-device model, E2B/4B, and answer length).

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

Table 4. Validation of the hypotheses.

Hypothesis	Result	Status
H1 Socratic	Posture 4.6/5 from *fine-tuning* (vs 1.6 ChatGPT, 1.2 raw Llama; n=40, κ/CI to be computed); 78 % better *perceived* retention.	Supported
H2 Trust	87 % trust vs 19 %; local processing perceived as reassuring.	Validated
H3 Offline	Local generation on all devices (30 to 7 tok/s); 85 % "no constraint".	Validated
H4 Orchestration	PoC validated (<800 ms) but paused in production (infrastructure).	Paused
H5 Portability	14 connectors (including 5 international) on a single canonical model.	Validated (archit.)
H6 Frugality	∼3 W; ∼0.005 Wh/request vs ∼0.24–0.30 Wh cloud (∼50–6 600× less).	Validated

Table 5. Compliance *by design* with European frameworks.

Framework	Requirement	Our app's ***by-design*** response
AI Act (education, high-risk)	Governance, human oversight, transparency	Local processing; neither grades nor steers; supervisable tutor
GDPR art. 25	*Privacy by design*, minimization	Learning data 100 % local; only non-nominative progress (XP) synchronized
GDPR art. 8 (minors)	Variable consent, 27 regimes	No personal school content transferred; legal basis of the account to be documented
US CLOUD Act vs GDPR	Risk of extraterritorial access	Inference and school content local ⇒ sensitive content not exposed
Green Deal	Digital sobriety	∼3 W, no datacenter

References

[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] Regulation (EU) 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*.