European Union Contest for Young Scientists

EUCYS 2026 · Kiel, Germany · Field: Computing

A Sovereign, Offline and Socratic AI Tutor

for cognitive autonomy and European digital sovereignty

Lucas Gerhardt Clément Bellet-Odent

Lycée Henri Matisse, Vence (France) · Supervising teacher: Pierre Aboussouan

Scientific report, French delegation

Our app is a school assistant built on artificial intelligence that runs **entirely on the student's device**. Where generative tools hand over the answer, feeding a « **cognitive debt** » measured by electroencephalography, our app adopts a strict **Socratic posture**: it never gives the solution, but guides through questioning. Powered by small open models (**Gemma**, quantized *by design* in int4 QAT), specialized through **QLoRA** *fine-tuning* and grounded by on-device **retrieval-augmented generation** (RAG), it turns the student's real lessons into quizzes, revision sheets and tutoring, **without their lessons or conversations ever leaving the phone**. This report establishes, evidence in hand, that a *private, frugal and sovereign* educational AI can, on the criterion that matters most for learning (**making the student reason**), do better than cloud models.

PARTNERS, SUPPORTERS & COLLABORATIONS

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

A sovereign, offline and Socratic AI tutor for cognitive autonomy and European digital sovereignty

Lucas Gerhardt Clément Bellet-Odent

Lycée Henri Matisse, Vence (France), Supervisor: P. Aboussouan · EUCYS 2026, Kiel · Field: Computing

Our app is **live on the App Store**, free and offline, adaptable to any curriculum and any language. The presentation video (2 min) is included in the submitted folder; this report is its scientific supplement and in-depth companion.

Abstract — Generative AI answers in the student's place and short-circuits the cognitive effort, an EEG study from MIT (2025) associates it with a collapse of brain connectivity (« cognitive debt »). Our app takes the problem the other way around: a *Socratic* tutor that never gives the answer but guides through questioning, run **100 % on the device**. It relies on small Gemma models **quantized by design** (int4 *QAT*, GGUF engine), specialized through QLoRA *fine-tuning* on a Socratic corpus validated by teachers (1 000 pairs, open license) and grounded by an on-device hybrid RAG. The student's learning data (lessons, conversations, grades) stay on the device, only non-nominative progression (XP, level) feeds the leaderboard : the architecture *answers by construction* to the GDPR and the AI Act and serves European digital sovereignty. A canonical data model projects **14 ENTs** (Pronote, ClasseViva, WebUntis, Moodle…) onto a shared pedagogical layer, making the tool portable to the **37.5 M secondary-school students of the EU** (∼87 M including the United States via Google Classroom). In blind evaluation (40 questions, 2 graders), *fine-tuning* installs a clear Socratic posture (**4.6/5**), absent from ChatGPT (1.6/5, which delivers the answer) as from the non-specialized base model (Llama 3.2, 1.2/5): it is specialization, not model power, that creates the posture, at ∼3 W and with no cloud transit. Six original contributions (**PAQ**, **PEDAL**, **ZED/W-Gap**, **ZPD-Steer**, **CDL/SRR**, **PS2**) move answer withholding from the *prompt* to *decoding*, *quantization* and *retrieval*, each reported under an honest status, with no fabricated result. *The best educational AI is not the most powerful one: it is the one that knows when not to answer.*

Keywords — on-device AI ; Socratic tutoring ; European digital sovereignty ; GDPR *by-design* ; quantization & QLoRA ; RAG ; frugality.

1 Introduction

In France, 60 % of students report poorly organizing their revisions (PISA 2022), and generative tools (ChatGPT, Photomath, Gemini) provide immediate answers that suppress the retrieval effort necessary for learning. The EEG study *Your Brain on ChatGPT* from the MIT Media Lab [9] ($N = 54$, 4 sessions) measures in LLM users the **lowest brain connectivity** (−55 % vs. the tool-free group) and observes that **83 %** of them are unable to quote a sentence they had just written (against ∼11 %). The researchers name this phenomenon the « **cognitive debt** ». Across the Union, 37.5 M secondary-school students[1] are exposed to it.

The opposite is nevertheless possible. A study from Harvard [7] ($N = 194$) shows that an AI configured as a **tutor** (brief, "one step at a time", without revealing the solution) doubles the learning gains compared to an active lecture. *The determining factor is not the power of the model, but its posture.*

[1] Eurostat, *Secondary education statistics*, 2024 (ISCED 2+3).

This issue has also become European and regulatory. The AI Act classifies education as « high-risk » (obligations applicable from August 2026), the GDPR mandates the minimization of minors' data, and the Commission's *Tech Sovereignty Package* (2026) recalls that the EU depends on non-European countries for more than **80 %** of its key digital technologies. An educational AI processing data *locally* answers these three requirements (pedagogical, legal and sovereign) in a single architectural move.

Research question. *How can one design an AI-based school assistant that strengthens the student's cognitive autonomy instead of substituting for it, while guaranteeing the privacy of minors' data and European digital sovereignty?*

Contributions. (i) A **Socratic** tutor constrained through *fine-tuning* and *system prompt* ; (ii) a **Local-First** architecture where all inference is on-device, embodying the GDPR and the AI Act *by-design* ; (iii) an **on-device hybrid RAG** (lexical + semantic) reducing hallucinations through grounding on local sources ; (iv) a **canonical data model** making the tool portable to 14 European school systems ; (v) a **blind evaluation** showing that the Socratic posture comes from specialization (our *fine-tuning*), not from model power (neither ChatGPT nor the raw Llama has it) ; and above all (vi) **six original contributions** (§3, Tab. 2), including **PAQ**, a posture-driven int4 quantization, a *research direction not yet measured*, which move the Socratic posture from the *prompt* to *decoding*, *quantization*, *retrieval* and *measurement*, entirely on the device.

2 Scientific background

Six bodies of work ground our design. The *cognitive load theory* [1] requires minimizing extraneous load and guiding « one step at a time ». *Spaced repetition* [2, 3] (meta-analysis: 839 measures) and the *testing effect* [4] justify flashcards and quizzes drawn from the real lesson. *Self-determination theory* [5] grounds a gamification of effort. The *assisted Socratic approach* [6, 7] shows that an AI constrained to question teaches better. Finally, *Edge AI* [11] makes possible a local processing of sensitive data. The active-learning paradox [8] (one learns more but *feels* learning less) justifies embracing Socratic friction.

These results converge towards a quantifiable requirement: reduce extraneous load (progressive guidance), force active retrieval (quizzes, flashcards at increasing intervals), support intrinsic motivation (gamification of effort), and process data locally.

Gap. To our knowledge, no solution combines *simultaneously*: strict Socratic posture, 100 % local privacy, gamification of effort, and pan-European portability/sovereignty. ChatGPT and Photomath hand over the answer (cloud, non-sovereign) ; cloud tutors of the Khanmigo type remain dependent on a connection and on third-party hosting ; raw open models run locally have neither pedagogical posture nor curricular grounding. Our app alone occupies the intersection (Tab. 1).

Caveat. This table is, by construction, partially subjective: other grids would favor each competitor (ChatGPT on raw quality, Khanmigo on the ecosystem). Our app does not claim to be « better everywhere » ; it claims to occupy *alone* the intersection of Socratic *and* 100 % local *and* sovereign.

3 Original contributions

Beyond its sovereign architecture and its measured maieutic posture, our app introduces contributions that move Socratic pedagogy out of the *prompt* layer, fragile and circumventable, and into **decoding**, **retrieval** and **measurement**, in an int4/GGUF regime on consumer hardware. Each is a *defensible composition of established primitives*, honestly labeled, implemented, prototype or proposed (pre-registered protocol): no figure is put forward without a non-fabricable

protocol. All share a single **local learner state**: *pedagogical rigor is enforced by construction, on the phone, not requested by instruction.*

Thesis: the Socratic posture *by construction*, at decoding

A maieutic *system prompt* is circumvented through social engineering. Our contributions make the posture **structural**: the decoding grammar renders the answer block *grammatically unreachable* (PEDAL, §5.6); the local gold answer becomes a *negative constraint* that is never exfiltrated (ZED, §5.7); the student's cognitive state, estimated on-device, drives retrieval, generation and scaffolding (ZPD-Steer, CDL, PS2); and **PAQ** (§5.3) allocates down to the *quantization bits* to preserve the non-answer. Pedagogy is no longer an instruction: it is a property of the system, all the way down to its weights.

4 Hypotheses and protocol

We treat our app as a research object through six testable hypotheses (Tab. 4): Socratic posture (**H1**), trust through locality (**H2**), offline operation (**H3**), local/cloud orchestration (**H4**), portability through a canonical model (**H5**), energy frugality (**H6**). Verification combines a **blind evaluation** (40 questions, 5 subjects, 2 graders), a **questionnaire** ($N = 47$ high-school students, Likert scale), **benchmarks** on 5 devices, and an **architecture audit**.

5 System: a local, sovereign educational AI

5.1 *Local-First* architecture

Our first prototype (V1) relied on the Gemini API: all conversations passed through third-party servers, compromising both privacy (**H2**) and offline use (**H3**). This realization (a "private" assistant that nonetheless sent every word of a minor to a non-European cloud) triggered the pivot to on-device. The guiding principle (Fig. 5) then becomes: school data is retrieved *directly* from the institution's server and stays on the device; AI inference runs entirely on it; no intermediate server is interposed. A cloud/local orchestrator (**H4**) was prototyped (switching <800 ms in the lab) then paused in the face of infrastructure limitations, in favor of a **100% local** operation, which, paradoxically, *reinforced* the sovereignty thesis.

The choice of on-device model results from a comparative evaluation under real-world conditions (iPhone 14, Samsung S23). Llama 3 (8B) was discarded (>15 s/response, >4 GB of RAM, unacceptable experience) in favor of a 4B model (median latency 2.7 s), dividing the response time by 5; a dynamic loading/unloading mechanism combined with 4-bit quantization brought the memory footprint below the threshold of devices with less than 4 GB.

5.2 On-device model and inference

The model runs on the phone's GPU via a **GGUF** engine (`llama.rn`/`llama.cpp`): 4 096 token context, quantized KV cache, *flash attention*, and a **grammar-constrained JSON generation** (GBNF) that makes quizzes and cards structurally valid. We adopt the open **Gemma** models in **int4 *quantization-aware training*** (QAT) [12], Gemma 3, and now **Gemma 4** (June 2026). QAT **integrates quantization *into* training** (target: the full-precision model), so that 4-bit retains a quality *nearly identical* to bf16: the 4B goes from **8 GB (bf16) to ∼2.6 GB in int4**, the perplexity degradation is **reduced by about 54%** compared to post-training quantization, and the smallest Gemma 4 (E2B) fits in **less than 1 GB**. A model **quantized *by design***, better *on-device* quality at a smaller footprint (→**H3**, **H6**).

5.3 PAQ, posture-aware quantization (*flagship contribution*, H1, H6)

Hypothesis (to be tested). We hypothesize that int4 quantization would **silently erode the Socratic posture**: under compression, the model would regress toward the reflex of *giving the answer*, even though the **perplexity would barely move**, an effect documented for alignment in general, but never studied from a pedagogical angle. *Mechanism.* The `llama.cpp` allocator (importance matrix, OBQ/GPTQ lineage) weights each weight column by its **activation energy** (mean of the squared activations) over a generic corpus, it therefore preserves what the model does *most often*: answer. PAQ would replace this criterion with a **contrastive behavioral margin** *Socratic-vs-answer*, derived from our leakage metric (W-Gap, §5.7), and would **route the int4 bit budget toward the weights that carry the non-answer**. From two sets of prompts built on a held-out subset of our validated Socratic corpus, P_{soc} (Socratic behavior) and its counterfactual P_{ans} ("give the answer directly"), we compute, per column i, the activation energy $A_{\text{soc},i}$ and a behavioral salience g_i, then a posture-importance score:

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

Injected into the importance matrix, s_i reorients the int4 quantization error toward the posture-bearing directions, *with tensor type and bit budget unchanged*: one would obtain an int4 GGUF of **identical size and latency**, whose error would be minimized along the directions that preserve the non-answer.

Novelty. To our knowledge, the first int4 bit allocation driven by a *contrastive behavioral margin*, formulating the Socratic non-answer as a *posture fragile to quantization*, distinct from AWQ (activation magnitude alone). **Status.** Prototype; no figure before execution. **Guardrail *(pre-registered).*** If the posture is *diffuse* (s_i not concentrated), PAQ reduces to the standard importance matrix and is reported *with no claimed gain.* **Validation.** Three equal-budget quantizations, B0 (no matrix), B1 (generic), B2 (PAQ), compared to fp16 on a held-out bench (W-Gap leakage, posture, perplexity, tok/s, energy). Criterion: *at equal size and latency, does PAQ recover toward fp16 a share of the posture that standard int4 loses?*

5.4 Specialization: dataset and QLoRA *fine-tuning*

Since generalist models ignore curriculum terminology, we specialize them on a Socratic corpus. A seed of ∼1 200 pairs written by teachers feeds a **knowledge distillation** (internal *Lucid Labs* platform); after **teacher validation** (*Lucid Professeur* interface), **1 000 pairs** retained out of ∼10 000 generated candidates (938 "gold", i.e. 93.8%) covering 18 subjects and 7 levels (6e–Tle) constitute the corpus, **released under an open license** (CC-BY-NC-SA 4.0). The *fine-tuning* uses QLoRA [13] (LoRA [14] on a frozen 4-bit NF4 model): the base model W_0 remains *frozen* and only two small low-rank matrices are trained, A (compression) and B (decompression), Fig. 2. Each layer then computes

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

With $r = 16$, only ∼0.1% of the parameters are modified ($r = 4$ destabilizes the posture, $r = 64$ overfits); the final validation loss $\mathcal{L} \approx 0.87$ stabilizes without train/val divergence (no manifest overfitting). A strict maieutic *system prompt* (never give the answer, identify the blockage, question in stages) reinforces the effect of the *fine-tuning* (→**H1**), as illustrated by the dialogue in Fig. 3.

5.5 Reliability: an on-device hybrid RAG

Fine-tuning alone left factual errors. We add an **entirely on-device retrieval-augmented generation (RAG)** [15] (Fig. 6), more robust than a simple vector search. Textbooks are split into sentence-bounded *chunks* (∼900 characters). For each question, two searches run on the device, **lexical BM25** (k_1=1.2, b=0.75, titles ×3) and **semantic** via **EmbeddingGemma** (256-dim vectors, Matryoshka truncation, L2 normalization), then are fused by ***Reciprocal Rank Fusion*** ($1/(60 + \text{rank})$). The best passages are injected into the context: the answers become **precise, local and sourced**.

5.6 PEDAL, answer withholding as a decoding invariant (H1)

PEDAL transforms the maieutic posture "never give the answer" from a *prompt* request, circumventable through social engineering, into a **decoding invariant**. *Layer 1 (structural)*: we compile a taxonomy of pedagogical moves (diagnosis, optional hint, mandatory question) into a GBNF grammar passed to `llama.cpp`/Gemma via the `grammar` field; no production allows a turn to end on a solution block, making the model *structurally* unable to conclude on a bare answer. A finite automaton (number of attempts, hint stage, persisted in WatermelonDB) selects the grammar variant at each turn, which also counters multi-turn posture collapse. *Layer 2 (lexical)*: for problems whose gold answer is available locally (quizzes, flashcards, RAG-grounded curriculum), we tokenize the canonical answer and its surface variants, then inject them as a hard ban (`logit_bias`), activable *lazily* after the diagnosis state (`grammar_lazy`/`grammar_triggers`) to leave the initial scaffolding free. Formally, at step t of decoding, the logits $\ell_t \in \mathbb{R}^{|V|}$ are corrected according to the state s_t of the automaton:

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

where $\mathcal{A}(s_t)$ is the set of tokens admitted by the GBNF grammar in state s_t (the term equals 0 if admitted, $-\infty$ otherwise) and $\mathcal{B}$ the set of tokens of the surface forms of the local gold answer: the answer block becomes of *zero probability by construction*. To our knowledge, this is the first system to enforce *simultaneously* the structure of Socratic moves and the literal withholding of the answer as a decoding constraint on a quantized phone LLM.

Novelty. A novel composition of established primitives (NeuroLogic-style lexical bans [19], decoding under GBNF grammar, taxonomies of pedagogical acts). **Status.** Prototype (Layer 1 immediate; Layer 2 on the subset with known answer). **Validation.** Pre-registered A/B/C study (prompt only / +grammar / +grammar+ban) on held-out problems and adversarial attacks; metrics: leakage, posture (blind grid, Cohen's κ), *alignment tax* (first-move correctness, tok/s, watts). A negative result remains publishable.

5.7 ZED, zero-egression decoding guard & W-Gap metric (H1, H2)

ZED requalifies answer withholding into a decoding invariant *and* provides its metric. *Level 1 (implemented, guaranteed)*: since the gold answer is local (RAG chunk, curriculum, generated solution), we tokenize its surface forms and ban them hard (`logit_bias`, confirmed in our `llama.rn` build), making the literal answer sequence *non-emittable* regardless of the *prompt*, a correctness guarantee robust to *jailbreak*, not an empirical score. *Level 2 (prototype)*: for *paraphrase* leaks that the lexical ban misses, a FUDGE-inspired discriminator [18] reuses EmbeddingGemma (256-dim) to score the proximity of the partial answer to the gold answer, through generation with restarted windows (`llama.rn` exposing no per-step logit hook, declared limitation, re-prefill cost

measured). The next token is resampled according to

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

where P_ϕ is the on-device logistic discriminator on EmbeddingGemma and $\eta > 0$ the guidance strength. The **W-Gap** leakage metric, deterministic and local, aggregates overlap and semantics *without a cloud judge*:

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

with g_n the n-grams, $e(\cdot)$ EmbeddingGemma and $\sum_k \lambda_k = 1$; a *hard leak* is declared if the exact final form or $\ge k$ contiguous gold tokens appear.

Novelty. The first answer-withholding guard at decoding *and* the first leakage metric (W-Gap) operating entirely on-device (int4/GGUF regime, without a cloud judge) for tutoring. **Status.** Level 1: implemented & measurable (correctness guarantee); Level 2 and W-Gap validation: prototype/preliminary. **Validation.** Adversarial-student protocol on three arms; the honest headline result is the W-Gap reduction on *paraphrased* leaks (the 0% literal leakage of Level 1 is a consistency check, not a discovery). W-Gap validity: correlation (Spearman/Pearson) with human judgments on 40 questions, hence *preliminary* ($N = 40$).

5.8 ZPD-Steer, the local learner state drives retrieval and generation (H1, H2)

Our app's revision engine (`services/vacation/elo.ts`) already maintains an *Elo-type* ability estimate, formally related to the Rasch model [20], ability θ per subject, difficulty λ per notion, updated online from the local revision log; but this signal is today confined to spaced repetition and invisible to the tutor, which adapts difficulty only through a static class-level chain. ZPD-Steer connects this learner state to the Socratic loop in two ways, *entirely on the device*. (1) **ZPD-Rerank**: a difficulty/readability score is precomputed per chunk at indexing, then a near-zero-cost term $\lambda \cdot f(d_{\text{chunk}} - \theta)$ is fused into the existing *Reciprocal Rank Fusion* block, favoring passages *slightly* above mastery (desirable difficulty in the sense of Bjork [22], Vygotsky's proximal zone [21]). (2) **ZPD-Decode**: θ, λ and the recently fragile notions are injected as structured state into the Socratic *prompt*, and the `targetDifficulty` function, already implemented, chooses the difficulty band of the next question. Personalization lives in the *local memory space*, not in the weights: without catastrophic forgetting, and instantly erasable. The ability θ comes from a Rasch model, the probability that a student of ability θ succeeds at an item of difficulty b is

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

θ being updated online from the local log, and the reranking of a chunk c for query q becomes

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

where d_c is the chunk difficulty and g a *desirable difficulty* kernel maximal at a small positive offset $\delta^\star$ (e.g. $g(u) = -(u - \delta^\star)^2$): passages *just* above θ are favored.

Novelty. To our knowledge, the first *on-device* implementation reranking a free-text corpus retrieval according to the fit between passage difficulty and the estimated ability of a *human learner*, adaptive tutors and *knowledge tracing* have estimated ability for a long time, but on the cloud side and outside free-text RAG retrieval; here the estimation is computed 100% on the device. **Status.** ZPD-Rerank: prototype (mechanism implemented, gain to be measured); closed-

loop decoding: prototype/preliminary. **Validation.** Pre-registered blind A/B (ZPD-Rerank *vs* relevance alone) on ∼150 tuples (query × simulated level), metric = level appropriateness; the difficulty labels are validated against teacher judgment *before* the A/B. Dataset released under an open license.

5.9 PS2, productive struggle scheduler by grammar switching (H1)

PS2 carries the idea of *wheel-spinning* and productive struggle, arising from step-structured intelligent tutors [23], into the free-text Socratic chat. From the local trace alone (attempts before success, time per step, error-then-self-correction sequences, repeated-identical *vs* varied errors measured by EmbeddingGemma similarity), a classifier estimates a **three-state struggle band** that *switches the move grammar variant* at decoding: in productive struggle, the grammar forbids hints and forces a question terminal; in frustration, it allows a hint stage (or even a worked example from *another* problem, leakage-safe); in under-stimulation, it skips the scaffolding. The band β is chosen by a classifier P_ψ over the signal vector $\mathbf{z}$ (attempts, time per step, inter-error similarity) and indexes the grammar variant $\mathcal{A}^{(\beta)}$ of eq. (3):

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

The switching uses `grammar_lazy`/`grammar_triggers`. PS2 thus operationalizes desirable difficulty in the LLM chat, which today has no accepted metric, and shares the single local learner-state object with CDL and ZPD-Steer.

Novelty. The first to close the loop "dialogue-derived struggle band → decoding-time switching of the Socratic move grammar", entirely on the device. **Status.** Switching controller: implemented & demonstrable; detector–teacher agreement: preliminary; learning gain: proposed. **Validation.** The controller is demonstrable today; the learning gain via a pre-registered randomized pre/post trial + delayed transfer (novel isomorphic items), metric = Hake normalized gain (g). Test data remaining on the device; no gain figure without a pre-declared N.

6 The application in practice

Our app is not a laboratory prototype: it is a **real mobile application**, live on the App Store and **downloaded by ∼1 000 students**. Fig. 7 shows four of its screens, all AI inference running **100 % on the device**, without any learning data leaving the phone.

7 Pan-European portability and sovereignty

7.1 A canonical data model

The European school landscape is fragmented: each country has its own systems (ENT) and its own curriculum. Rather than one application per country, we project each ENT, through **a single adapter**, onto a **shared canonical model** (`services/shared`: *Grade, Homework, Timetable, News, Attendance, Chat…*), on top of which the entire pedagogical layer is written *only once* (Fig. 9). The marginal cost of a new country then reduces to two building blocks: an ENT adapter and a curricular specialization (dataset + RAG in the target language). Our app already ships **14 connectors**, including 5 international ones: ClasseViva (Italy, >2.5 M users/day), WebUntis (German-speaking area, >26 000 institutions), Moodle (>500 M users), Canvas and Google Classroom; Pronote alone covers the French market (18 M users, SecNumCloud 3.2). At the heart of the European target, **37.5 M secondary-school students** in the EU27, the **United States** market is added through **Google Classroom** (∼49.5 M K-12 students)[2], i.e. **∼87 M**

[2]Eurostat 2024 (EU, ISCED 2+3); NCES 2024 (K-12 USA); Google for Education, 2025.

students (EU + USA) and **150 M users** of Google Classroom worldwide. Without any new infrastructure, our app plugs into what already exists (Fig. 8).

7.2 European compliance *by design*

The AI Act [16] classifies education as "high-risk" (Annex III.3), with obligations applicable from 2 August 2026. Local processing of inference natively satisfies data governance, minimization and human oversight; since education remains "high-risk" by default, our app's maieutics (which neither grades nor steers) falls under the derogation of art. 6(3), a narrow procedural task, without substitution for teacher judgement, which remains to be confirmed. On the GDPR side (art. 25, *privacy by design*), **learning data (courses, conversations, grades) do not leave the device**; only non-nominative progress (XP, level) is synchronized for the leaderboard, with no personal school content exposed to the *US CLOUD Act*. The stake is also strategic: the Draghi report (2024) notes that only **4 of the 50** top technology companies worldwide are European, and the *EuroStack* initiative (∼300 Bn€) advocates an "*open source first*" principle. Where Gaia-X illustrates the limits of "top-down" sovereignty (bureaucracy, capture by hyperscalers), our app embodies a **"bottom-up"** sovereignty, distributed across millions of devices *already* owned and built on open building blocks (Gemma, `llama.cpp`/GGUF, EmbeddingGemma, open dataset): no datacenter to build, sovereignty fits in the student's pocket (Tab. 5).

8 Results

8.1 Speed and frugality (H3, H6)

Local generation is operational on all tested devices (Tab. 3), from 30 tok/s on iPhone 14 to 7 tok/s on the Lenovo tablets distributed by the Région Sud, slow but usable for tutoring where the student reads each answer. Above all, inference is **frugal**: ∼3 W, i.e. ∼0.005 Wh/request (≈3 W × 6 s) against ∼0.24–0.30 Wh for a cloud request [17], and up to ∼33 Wh for a *reasoning* model (Fig. 12), i.e. ∼50 to ∼6 600 times less (usage energy; manufacturing and recharging of devices, already owned, not counted). The context makes the gap decisive: datacenters consumed ∼415 TWh in 2024 (∼1.5 % of global electricity), ∼945 TWh projected in 2030 (IEA), and a 100-word email written by ChatGPT would mobilize ∼519 mL of water (UC Riverside, 2025). Consumption **measured with a wattmeter** during continuous generation (Fig. 10).

8.2 Pedagogical quality (H1)

On 40 questions graded blind by 2 graders (Fig. 13), ChatGPT excels, as expected, in relevance (4.5) and rigor (4.2); but designed to deliver the answer, it fails on pedagogical posture (**1.6**) and on curriculum adaptation (2.9). Tellingly, the raw base model (Llama 3.2) also obtains a weak posture (**1.2**) that our *fine-tuning* raises to **4.6**: it is specialization, not model power, that installs the posture. On the **four pedagogical criteria**, our app scores **4.2/5 against 3.3** for ChatGPT (the structural *offline* criterion, 0 vs 5, already covered by **H3**, is treated separately). These gaps remain a *robust trend* (κ/CI to be computed), not a definitive proof.

8.3 Trust and adoption (H2)

In the survey ($N = 47$), **89 %** would recommend our app, **74 %** prefer it to ChatGPT ("it forces you to think"), **78 %** report better retention. On privacy, the gap is clear: **87 %** trust our app (4.4/5) against **19 %** for ChatGPT (2.3/5), local processing is understood as a concrete advantage. Six months after launch, the application has been **downloaded by ∼1 000 students** (including ∼250 active users) (Fig. 11). Tab. 4 summarizes the validation of the hypotheses.

8.4 CDL, an on-device proxy for "cognitive debt" (H1)

The *Cognitive Debt Ledger* computes, from the conversation trace already local (WatermelonDB) and the engine's token counts alone, a per-turn behavioral vector, *without any sensor or data leaving the device*. Central signal: the **Student Reasoning Ratio** at turn τ,

$$\mathrm{SRR}_\tau = \frac{n^{\text{reas.}}_{\text{student},\tau}}{n^{\text{reas.}}_{\text{student},\tau} + n^{\text{reas.}}_{\text{tutor},\tau}}, \tag{9}$$

where $n^{\text{reas.}}$ counts reasoning tokens, complemented by the answer-request rate $\rho_{\text{req.}}$, the attempt-before-hint rate $\rho_{\text{att.}}$ and the latency before first attempt. Aggregated over the session, these signals form a *cognitive debt index*

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

The separation of reasoning tokens *reuses PEDAL's GBNF move grammar* (eq. (3)) which already segments a tutor turn into {diagnostic | hint | question}; the student's turns are classified by a lightweight logistic head over the already-resident EmbeddingGemma embeddings, without downloading any model. A sustained drop in SRR makes the tutor *escalate* toward questioning, closing the loop that *benchmarks* merely measure. We do not claim that the CDL *measures* cognitive debt: it is a *behavioral proxy*, distinct from the neural construct measured by EEG [9]. **Novelty.** First fully on-device proxy of *cognitive offloading* (no egress), grounded in the move grammar to isolate the tutor's reasoning tokens. **Status.** Proposed (designed, not validated); labeling reliability: measurable; SRR↔offloading correlation: preliminary. **Validation.** Pre-registered construct-validity study on the $N = 47$ cohort: correlation (Pearson & Spearman) between SRR and self-reported offloading items, *conditioned* on an acceptable inter-annotator κ. With $N = 47$ being small, a null result is a likely and honest outcome.

9 Impact, limitations and ethics

Impact. Our app addresses three European divides. *Equity*: with no subscription and no connection, on a 150 € tablet, it reaches the students the cloud leaves behind (poorly connected areas, modest backgrounds). *Ecology*: at the scale of 37.5 M secondary-school students, ∼10 interactions per class day represent $\approx 6.8 \times 10^{10}$ requests/year. By shifting these requests from the cloud (∼0.3 Wh; ∼0.3 mL of water, a conservative estimate per short request) to on-device, the following is avoided each year, as an order of magnitude:

$$\underbrace{6.8 \times 10^{10} \times 0.3\,\mathrm{Wh}}_{\approx 20\ \mathrm{GWh}} \quad \text{and} \quad \underbrace{6.8 \times 10^{10} \times 0.3\,\mathrm{mL}}_{\approx 20\,000\ \mathrm{m^3\ of\ water}},$$

i.e. the electricity consumption of ∼5 000 European households and the volume of ∼8 Olympic swimming pools. *Sovereignty*: an AI that belongs to the student and stays in Europe. These axes align with UN SDGs 4, 10, 12 and 13 (Fig. 14). By running on older devices, our app also extends their lifespan, a material frugality that adds to its energy frugality.

Ethics. Building an AI for minors demands heightened vigilance: model biases (mitigated by teacher validation and RAG grounding on official textbooks), the risk of a *new* dependence (countered *by design* through answer withholding and Socratic friction), and transparency (dataset released under an open license, a patent application on the architecture filed with INPI). No commercial purpose guided the scientific design.

Limitations and biases. Small sample (47 students, one institution; limited external validity, a multi-institution study with $n > 200$ is planned); selection bias (volunteers); one grader is our supervisor (an evaluation with a broadened jury is planned); models >4B incompatible with <3 GB of RAM; full curriculum specialization for France only, with other countries belonging to the roadmap. The reported "perceived retention" is not an objective measure of learning and calls for a controlled pre/post-test protocol.

Honesty. The MIT study we cite has a modest sample size and a debated EEG methodology [10]: we draw on it as *motivation*, not as proof. Our released dataset comprises 1 000 *validated* pairs (938 "gold"), selected from ∼10 000 generated candidates. RAG reduces hallucinations without eliminating them; the Socratic posture, which questions rather than asserts, mechanically limits their reach.

10 Conclusion

Our app demonstrates, with supporting measurements, that an educational AI can *amplify* reasoning instead of substituting for it, while remaining private (87 % trust vs 19 %), frugal (∼3 W) and sovereign, and portable across a continent via a canonical data model. Outlook: completing the local⇄cloud orchestrator, multimodal tutoring (vision, voice), extending the dataset/RAG to other European curricula and languages, and an open accuracy/energy benchmark for on-device educational AI. *The best AI is not the one that gives the answer, it is the one that helps to find it.*

Partner ecosystem and scaling up. This work draws on an ecosystem spanning the entire chain: **Polytech Nice Sophia** (Université Côte d'Azur) guided us on the hybrid architecture and on-device optimization; the **teachers of Lycée Henri Matisse** designed and validated the corpus; **AxePI** protected the innovation (INPI patent) and hosted the server; **Lucarne Pro** handled publication on the App Store; the **Fondation CGénial** and **Sciences à l'École** supervised and qualified the project for EUCYS.

> *"At a time when so many students simply ask an AI for the answer, these ones did the opposite: they wanted to put the human back into the project, to build it with us, the teachers, and to always look further, instead of remaining in this 'AI generation' that no longer thinks."*
>
> , Carolle Chabanier, mathematics teacher, Lycée Henri Matisse

Recognition, Prix Orange Numérique

Our app received the **Prix Orange Numérique** (Concours Science Factor 2026). This support paves the way for moving from prototype to larger-scale deployment, within a logic of responsible, inclusive and sovereign educational technology.

Acknowledgements. Our thanks to Mr. P. Aboussouan, to all of our partners listed above, to the student testers and to our families. Our app is a **national laureate of the Concours C'Génial 2026** (EUCYS qualification) and a **double laureate of the Concours Science Factor 2026**, and is live on the App Store (∼1 000 downloads).

Supplement (separate document): detailed evaluation protocol (rubric, bias control, statistical treatment) and construction of the Lucid-Socratic-EDU corpus.