Scientific summary of the project

A sovereign, offline and Socratic artificial-intelligence 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

---

The problem. In France, 60 % of students report poorly organising their revisions, and generative AI tools (ChatGPT, Photomath, Gemini) provide immediate answers that remove the retrieval effort that learning requires. An electroencephalography study from MIT (2025) measures, in users of these tools, a marked drop in brain connectivity, a phenomenon the researchers name "cognitive debt". Conversely, a Harvard study shows that an AI configured as a tutor, brief and without revealing the solution, doubles learning gains compared with an active lecture. The determining factor is therefore not the power of the model, but its *posture*.

Our solution. We present a school assistant built on artificial intelligence that runs **entirely on the student's device**, with no connection. Rather than handing over the answer, it adopts a strict **Socratic posture**: it never gives the solution, but guides the student through questioning, one step at a time. From the student's real lessons, it generates quizzes, revision sheets and tutoring sessions, while keeping the learning data (lessons, conversations, grades) on the device.

The technology. The assistant relies on the small open **Gemma** models from Google, *quantized by design* to int4 through *quantization-aware training* (QAT): quantization compresses the model so that it fits on a phone (a 4-billion-parameter model goes from 8 GB to about 2.6 GB), and by integrating this compression into the training process, QAT preserves a quality nearly identical to full precision. Inference runs on the phone's graphics processor through a **GGUF** engine. A **QLoRA** *fine-tuning*, carried out on a Socratic corpus of 1 000 teacher-validated pairs, instils the maieutic behaviour by training only two small low-rank matrices (about 0.1 % of the parameters), which makes the specialisation light and reversible. For reliability, an entirely on-device **retrieval-augmented generation (RAG)** combines a lexical search (BM25) and a semantic search (EmbeddingGemma), fused by *Reciprocal Rank Fusion*, in order to ground each answer in the student's real textbooks and to reduce hallucinations. A **decoding grammar** (GBNF) enforces the structure of the Socratic "moves" and renders the literal answer grammatically unreachable, moving the pedagogical posture from the *prompt* layer (circumventable) to the *decoding* layer (structural). Finally, a **canonical data model** projects each school digital workspace (ENT) onto a shared layer through a single adapter, on top of which all the pedagogical intelligence is written only once; several connectors are already implemented (Pronote, ClasseViva, WebUntis, Moodle, Canvas), making the tool portable at the scale of the **37.5 million secondary-school students** of the European Union. Six original research contributions (PAQ, PEDAL, ZED/W-Gap, ZPD-Steer, CDL/SRR, PS2) extend this architecture by enforcing answer withholding as an invariant of *decoding*, *quantization* and *retrieval*, and by estimating the student's cognitive state entirely on the device.

Sovereignty and privacy. Because inference and learning data do not leave the device, the architecture answers *by construction* to the GDPR and the AI Act, and serves European digital sovereignty: no datacenter to build, no school content transferred to a non-European cloud. Only non-nominative progress (XP, level) is synchronised to feed the gamification leaderboard.

Results and impact. 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. **87 %** of surveyed students trust the app with their data, against 19 % for ChatGPT. Inference is **frugal** (about 3 W, the order of magnitude of an LED bulb), with no cloud transit and no cooling water. The project thus acts on three European divides: *equity* (a free, offline tutor on a modest device that reaches the students the cloud leaves behind), *ecology* (a frugal AI that extends the life of older devices) and *sovereignty* (an AI that belongs to the student and stays in Europe). *The best educational AI is not the most powerful one: it is the one that knows when not to answer.*