Learning in the Age of AI:
A Sovereign, Metacognitive Approach to Higher Education

Alexandre Amrani, Emma Choukroun, Gautier Miralles

UTC EduTech

Université de Technologie de Compiègne (UTC)

Challenge: AI Grand Challenge 2026 — Inria

I. The Situation: AI, Learning, and the Challenge of Cognitive Engagement

Higher education finds itself in a predicament that is without recent precedent. The very technology that promises to democratise access to knowledge also threatens to render obsolete the mechanisms through which that knowledge was traditionally acquired. Since the public release of large language models in late 2022, students across the world have gained access to systems that can generate essays, solve problems, and explain concepts at a level that rivals or exceeds their own capabilities. The information that once required hours in the library, then minutes on a search engine, now arrives in seconds, complete, fluent, and often indistinguishable from work produced by a human hand. What was once scarce is now abundant, and the advantage that teaching drew from its position as the gatekeeper of knowledge has, as Pr. Youssef Gahi describes it in his white paper on higher education in the age of AI assistants [7], largely dissolved.

This situation did not emerge from nowhere. It is the product of a longer arc that began with the massification of higher education, with global enrolment more than doubling since 2000 to reach over 264 million students according to UNESCO, and the corresponding strain on institutional resources. Technology was initially offered as the solution: online learning platforms, MOOCs, and adaptive systems promised to deliver high quality education at lower cost, yet the results have been decidedly mixed, with improved access but limited evidence of deeper learning outcomes.

Into this strained ecosystem, generative AI has arrived with the force of an external shock. Alain Goudey, Associate Dean for Digital at NEOMA Business School, presented during the AI Grand Challenge keynote series data from Anthropic's survey of over 80,000 users across 159 countries showing that Western Europe exhibits the highest global AI skepticism at 35.6 percent negative sentiment, while simultaneously demonstrating that European users index double the global average on privacy concerns [8]. This is not, Goudey argued, a cultural handicap but rather a sophisticated design specification: European skepticism functions as a high bar quality standard, a demand for AI that is sustainable, privacy respecting, and genuinely useful rather than spectacularly disruptive. His research, published in the *Communications of the Association for Information Systems* with Régis Meissonier and colleagues [9], analysed the legitimacy of generative AI in French higher education across the pragmatic, cognitive, and moral dimensions, revealing that students and teachers inhabit structurally different legitimacy profiles, with students more focused on short term pragmatic advantages while teachers foreground the risks of metacognitive laziness and academic dishonesty.

The problem, in other words, is not that students have access to AI. It is that the dominant mode of access, the frictionless, answer giving interface of general purpose chatbots, systematically bypasses the cognitive processes that produce durable learning. Research stretching back decades, from Roediger and Karpicke's work on the testing effect [17] to Kapur's paradigm of productive failure [11], converges on a finding that should give every educator pause: learning that feels effortful produces more durable long term retention than frictionless delivery. The desirable difficulties framework, formalised by Bjork and Bjork [2], establishes that the very struggle that students seek to avoid is the mechanism through which understanding is built. When that struggle is removed, what remains is not efficient learning but its simulation, the completion of tasks without the construction of competence.

This is not merely a pedagogical preference. It is now empirically documented. Stadler, Bannert, and Sailer showed in a 2024 study published in *Computers in Human Behavior* that students who outsourced reasoning to large language models showed significantly lower conceptual understanding on delayed tests, even when they performed identically on immediate tasks [18]. The cognitive atrophy that Goudey identified as one of five key risks, and that he proposed to counter through deliberate habits including writing one's intent before one's prompt and auditing one's own atrophy regularly, is not a speculative concern but a measurable phenomenon.

The situation, then, is this: universities are operating in a context where the traditional mechanisms of learning are being silently bypassed at scale, where students and faculty have divergent perceptions of what is at stake, and where the default technological response, i.e. more AI providing more frictionless answers, may be making the problem worse. The expert resources provided by the challenge organisation reinforce this diagnosis. A large-scale synthesis of existing studies by Di Pietro and Castaño Muñoz, published in *Computers & Education* (2025), examines the effect of educational technology on less advantaged students and finds that the benefits of digital tools are far from guaranteed, depending instead on implementation quality and pedagogical design [5]. The white paper by Pr. Gahi identifies the phenomenon of the student who receives a task, passes it to an AI, copies the result, and submits it, producing a deliverable without ever having constructed the underlying competence [7]. And Arnaud Lévy's analysis of the failure of educational technology offers a broader structural critique: the wholesale adoption of digital technology in education has, by his account, largely failed to deliver on its promises, a conclusion that demands we examine not just the technology but the assumptions embedded in its design [13].

II. Analysis: Why Current Educational AI Falls Short

The diagnosis above points toward a paradox: the more AI is deployed in education, the less learning may be occurring, at least of the kind that education is supposed to produce. But this paradox deserves careful unpacking, for it is not inherent to AI itself but to the specific design choices that characterise most current implementations.

The most visible failure mode is the system that helps students complete assignments in ways that undermine the learning those assignments were designed to produce. When a student can paste a problem statement into a chatbot and receive a fully worked solution in seconds, the incentive to struggle, to make mistakes, to persist through confusion, and to experience the satisfaction of eventual understanding is replaced by the incentive to optimise for output. This is not cheating in the traditional sense; it is a rational response to an environment in which the effort of thinking has been made avoidable. The student who takes the shortcut is not necessarily lazy or dishonest; they are behaving rationally within a system that has not yet adapted to the new technological reality.

A second approach has been to restrict access: blocking AI tools on institutional networks, designing assessments that assume an absence of AI, or deploying AI detection software. This strategy is equally fragile, not least because students already access powerful models from their personal devices, beyond institutional control. Yet there is a deeper problem: restriction frames the issue as one of rule enforcement rather than pedagogical design. It assumes that the goal is to catch students who circumvent the system rather than to create conditions under which circumventing the system is no longer the rational choice. The EU AI Act, as Alexandra Prégent from CEA Saclay noted in her keynote on AI ethics [15], classifies AI systems used in education as high risk, but the regulatory response, while necessary, cannot substitute for pedagogical imagination.

The third approach, and the one most prevalent among institutions that have moved beyond the initial panic, is the deployment of educational AI that is essentially a rebranded version of the same general purpose technology: an AI system that explains concepts, answers questions, and generates practice problems. While this represents an improvement over a raw chatbot, it inherits a fundamental limitation: it operates on a logic of content delivery. Its goal is to transfer information from the model to the student, adjusted for difficulty perhaps, but not fundamentally different from a textbook that talks back. The evidence on intelligent tutoring systems, as synthesised by Kulik and Fletcher [12] and VanLehn [19], shows that these systems can produce genuine learning gains, with effect sizes of 0.66 to 0.79 standard deviations. Yet the most successful systems (from Carnegie Mellon's Cognitive Tutors to the Open Learning Initiative that Justine Cassell recommended in her exchange with our team) achieve their results through granular modelling of student knowledge and carefully structured feedback, not through the open ended conversation that current LLMs enable.

What the current wave of educational AI has largely failed to address is the metacognitive dimension of learning, that is the capacity of students to monitor their own understanding, to identify what they do not know, to select appropriate strategies, and to adjust their approach when something is not working. Self regulated learning, as Zimmerman formalised [21], is one of the strongest predictors of academic success. Yet most AI systems interact with students as if learning were a passive process of absorption rather than an active process of construction. They treat the student as a vessel to be filled rather than a thinker to be provoked.

This is where our analysis converges with multiple keynote sessions, each of which illuminated a different dimension of the failure. Caroline Beslin, project methodology expert at École Centrale de Lyon, presented her framework arguing that any project must be consolidated across four dimensions (Intellectual, Material, Social, and Emotional) and that technology teams systematically develop the first two while neglecting the latter [1]. The Social and Emotional dimensions, she argued, are the bridge between technical depth and institutional legitimacy; decision makers who determine whether a project succeeds are most often not technical experts, and they need to understand a project immediately. Beslin's framework reveals that the current failures of educational AI are not primarily technical failures; they are failures of social and emotional design. The systems do not understand who the student is, what they fear, what they aspire to, or what would make them feel supported rather than surveilled.

The tension that Prégent raised around the distance between what a development team knows and what would be required for a complete ethical assessment applies directly here [15]. The teams building educational AI are predominantly engineers and computer scientists; they are not learning scientists, not developmental psychologists, not classroom educators, and most importantly, not students. The harm foreseeability problem that Prégent described (the fact that harm is always standpoint dependent) means that the very design choices that seem optimal from an engineering perspective may be harmful from a pedagogical or developmental one.

The AI that explains everything may seem helpful; the AI that refuses to explain may seem obstructive. The research on desirable difficulties suggests that the opposite may be true.

Bastien Guerry, former developer at DINUM and open source maintainer, introduced a perspective that is no less incisive in his keynote [10]. His analysis of what open source actually means for AI (whether the weights are the equivalent of source code, whether a model with usage restrictions can genuinely be called open, whether new legal categories are needed) raises direct questions for educational AI. If a university deploys an AI system that uses a model whose license restricts its use or whose training data cannot be audited, the institution has not achieved sovereignty over its educational technology; it has merely outsourced pedagogy to a vendor with more restrictive terms than traditional textbook publishing. Guerry's historical analysis of how software got protected, through the 1980s shift toward aggressive software patenting, serves as a cautionary tale for educational institutions that may be sleepwalking into dependency on proprietary AI infrastructure.

What emerges from this analysis is a clear gap. Current educational AI is strong on content delivery but weak on cognitive development. It can answer questions but cannot diagnose why a student is asking them. It can generate exercises but cannot track which misconceptions a student is most prone to. It can converse naturally but cannot distinguish between a student who is genuinely stuck and one who is trying to get the answer by pretending to be stuck. And the infrastructure on which it runs, i.e. the cloud APIs of American technology companies, embeds a set of assumptions about data ownership, privacy, and pedagogical control that may be fundamentally at odds with the values of European higher education.

III. Positions Debated and Arbitrated

No serious proposal emerges from a single line of reasoning. Our team's thinking evolved through internal debate, external challenge from the AI Grand Challenge community, and the provocation of the keynote sessions. What follows is an account of the positions we weighed and the arbitrations that shaped our final contribution.

The most persistent internal debate concerned the relationship between cognitive friction and equity. A model that requires a student to articulate their current understanding before receiving assistance imposes a cost: it privileges students who can express themselves clearly in writing and penalises those who cannot. The same reasoning trace that produces deep learning for a fluent writer may produce frustration and disengagement for a student with lower written fluency, a student who is not a native speaker of the language of instruction, or a student whose cognitive style is more visual than verbal. A possibility would be to accept this equity cost on the grounds that the alternative (as an answer giving system that bypasses cognitive effort) produces even worse outcomes for disadvantaged students, who are the most reliant on education as a route to opportunity. A system that gives answers without building understanding, under this logic, deepens existing inequalities by credentialing students without equipping them. But what would it mean for the students who are already marginalised by our educational institutions? In regards to the research on desirable difficulties, we recognised that the relationship between friction and learning is not linear. The zone of proximal development, as Vygotsky formulated it [20], provides the theoretical anchor: the optimal challenge level varies across students, and a well designed system must calibrate to that level dynamically, not impose a uniform standard.

Mahdi Ayadi's comment on our proposal pushed us further on this point [22]. He argued that students optimise for ease, and that a tool designed for healthy learning will not stop them from turning to an easier alternative for an effortless answer. His team's research, showed that students know they are acting counterproductively and do it anyway, which means the problem is not informational but behavioural. This reinforced our

conviction that the response must be structural rather than voluntarist: the requirement to think before receiving help cannot be a toggleable option but must be the default interaction pattern, precisely because students will rationally choose the easier path when one is available. Our resolution was therefore to design for multiple style of reasoning expression. This is not a compromise but a principled position: equity and cognitive rigour are not trade offs to be balanced but design parameters to be jointly optimised.

A second debate concerned the tension between sovereignty and model capability. Our commitment to running open source models on institutional servers rather than relying on commercial cloud APIs carries a performance cost. The most capable models currently available are proprietary and cloud hosted, while open source alternatives such as Mistral and Llama are improving rapidly but still lag on complex reasoning tasks, nuanced pedagogical judgment, and multilingual fluency. A possibility would be to prioritise pedagogical quality over architectural purity: if a proprietary model delivers superior Socratic tutoring, then that is the tool students deserve, and the sovereignty objective can be pursued through data governance rather than infrastructure control. In regards to Guerry's analysis of the legal status of AI weights [10], however, a model that is merely accessible rather than genuinely open does not confer sovereignty. Moreover, as Goudey argued [8], European scepticism of AI is not a bug to be worked around but a feature to be designed for, a market signal that European users want systems that respect their privacy and cognitive autonomy, even at the cost of raw capability.

The arbitration of this debate was both practical and principled. Practically, our experiments with the infrastructure provided by our university (UTC), which enables inference on open source models including Mistral and Gemma families on university servers, demonstrated that when the system prompt is carefully engineered with Socratic constraints, student state classification, and profile conditioned context, a model of modest size operating on institutional hardware can produce pedagogically effective interactions. The research by Favero et al. [6], which achieved statistically significant Socratic behaviour with a fine tuned smaller model, provided independent validation of the feasibility frontier. On principle, we recognised that a system which cannot be audited, modified, or deployed independently is not a system that an educational institution genuinely controls, and that the loss of control over the pedagogical layer is a more serious risk than the loss of some percentage points on standard benchmarks.

The third major debate revolved around the question of paternalism. Is it legitimate for an AI system to refuse a direct request (a student asking for the answer without effort) on the grounds that compliance would harm the student's learning? A possibility would be to argue that the student is an adult, capable of making their own decisions about how to use a tool, and that a system which overrides their expressed preference violates their autonomy. Self determination theory, as developed by Deci and Ryan [4], suggests that autonomy is a basic psychological need and that a system undermining it will, in the long run, weaken intrinsic motivation. But what would it mean for a system to remain neutral in the face of behaviour that undermines the very purpose of education? Educators routinely make decisions that students do not like (assigning difficult problems, withholding answers, insisting on revision) precisely because those decisions serve the student's long term learning interests. An AI that does not exercise similar pedagogical judgment, under this logic, is not neutral but negligent; its refusal to refuse makes it complicit in the student's self defeat.

In regards to Prégent's distinction between prevention and post assessment in ethical AI design [15], the question is not whether the system should ever decline a student request, but under what conditions and with what transparency. Our resolution was to design a graduated response structure: the system does not simply refuse but acknowledges the student's frustration, names the pedagogical logic transparently, and offers a

structured choice between alternative forms of help. The firmness escalates only when the pattern of seeking shortcuts repeats, and even then, the response is designed to feel like reengagement rather than punishment. This is the approach we implemented in our demonstration of what happens when a student insists on a direct answer: the system pivots in a way that makes the thinking path feel like the faster option. However, we could not only propose this alternative. As other AI chatbots are already used by students and give direct answer, we implemented an option to send messages to a direct chatbot with no socratic discourse to prevent the desire to go to another platform.

A fourth line of debate concerned the learning profile itself: what it should track, how it should be updated, and who should control it. Chinmay Das raised a prescient question in the public commentary: how would the learning profile be designed and updated over time to avoid reinforcing incorrect assumptions about a student's abilities or learning style [23]? This question cuts to the heart of the personalisation challenge. A profile that cannot adapt risks pigeonholing students into static categories; one that adapts too readily may oscillate unhelpfully; one that adapts based on the wrong signals may reinforce the very misconceptions it is meant to correct. A possibility would be to make the profile primarily student controlled, with the student deciding what the system knows about them, editing it freely, and having the final say over its contents. In regards to the purposes of education, however, students do not always have accurate self knowledge, and one of the goals of education is precisely to reveal blind spots in one's own understanding. A profile limited to what a student chooses to disclose would be of limited pedagogical value.

Our resolution was a hybrid model in which the AI generates inferences that are presented transparently to the student in its User Profile, who can then annotate them. This draws on the distinction between episodic and semantic memory from cognitive science: the profile stores abstracted inferences about reasoning style, recurring misconceptions, and preferred formats rather than raw interaction logs, and these inferences are made visible and editable. The profile update mechanism, which we implemented as a structured process running every few messages, generates output that the student can review. This does not fully resolve the tension, and Chinmay's concern about reinforcing incorrect assumptions remains valid, but it shifts the dynamic from a system that silently categorises a student to one that makes its categorisations legible and negotiable.

Finally, we engaged deeply with the challenge posed by Justine Cassell's intervention [25]. Noting that the notion of a learning profile underlies most successful AI tutors developed over the last fifteen years, she specifically recommended examining the Open Learning Initiative from Carnegie Mellon University. Her point was not that our idea was unoriginal but that we stood on the shoulders of a substantial research tradition and needed to engage with it seriously. This led us to investigate OLI's architecture in depth. OLI's approach — embedding interactive, question based tutoring directly into reading material to check understanding in real time — shares our commitment to active engagement over passive reception. What OLI does not do, however, is build persistent individual learning profiles that travel with the student across courses and institutions, nor does it operate on infrastructure that gives students control over their data. Cassell's challenge strengthened our proposal by forcing us to articulate not just what we were adding but what was already there and how we differed from it.

IV. Contributions: A Learning Companion for Reflective, Sovereign Higher Education

The analysis above leads toward a specific set of contributions. They are not presented as a finished solution, rather an MVP with a coherent set of design principles, a conceptual architecture, and the conditions under which such a system could realistically operate.

What is the underlying philosophy? The answer is what we call a model that prioritises reflection before delivering answers. Instead of the dominant paradigm in which a student asks and the AI answers, we propose an interaction pattern in which the AI engages the student in metacognitive reflection before providing any substantive assistance. The interaction follows a structured sequence: the student externalises their current understanding, the AI diagnoses what is missing or mistaken, the AI delivers a precisely targeted intervention that addresses only what the student actually needs, and the student consolidates by reexplaining in their own terms. This is not a refusal to help; it is a refusal to help in a manner that bypasses thinking. The answer comes, but it comes after the student has done something with their mind. This principle is grounded in the cognitive science of learning (retrieval practice [17], productive failure [11], self explanation [3]) and operationalised through the prompting and context engineering techniques we developed. The system prompt architecture follows a layered design: identity and absolute rules at the base, pedagogical mode and student state classification in the middle, learner profile and course context as dynamic variables. This design was directly inspired by the StratL framework [16] and by the lessons learned from public accounts of earlier educational AI systems about the risks of static friction and context blindness.

The second contribution is the metacognitive student learning profile: a structured, portable representation of the learner's cognitive characteristics, including their reasoning style, recurring misconceptions, demonstrated competencies, preferred interaction formats, and metacognitive tendencies. The profile is not a static label but a dynamic model updated through each interaction, with the student retaining full ownership and control over its contents. It serves three functions simultaneously. For the student, it functions as a cognitive mirror, making visible patterns of thinking and learning that might otherwise remain invisible, and supporting the development of self regulated learning skills. For the AI, it functions as a personalisation substrate enabling the system to target exercises, explanations, and interactions to the student's specific zone of proximal development rather than delivering generic content. For the institution, it functions as an anonymised pedagogical intelligence layer revealing systemic patterns of misconception, curriculum gaps, and effective teaching strategies without requiring access to individual student data. The design of the profile schema draws on Bloom's taxonomy for cognitive level classification, on Zimmerman's model of self regulated learning [21] for the metacognitive dimensions, and on our research synthesis on learner profile context engineering that identified seven key dimensions: reasoning style, recurring misconceptions, preferred exercise format, current difficulty level, indicators of being stuck, engagement signals, and recent progress.

The third contribution concerns the infrastructure on which such a system operates. We argue that pedagogical sovereignty and technical sovereignty are inseparable: an educational AI that depends on proprietary cloud infrastructure for its core reasoning cannot genuinely serve the values of public higher education, because the terms of that dependency ultimately determine what the system can and cannot do. Our proposal is for a system that operates on university infrastructure running open source models. This is not a technocratic preference but a pedagogical one, for three reasons. First, it ensures continuity of access: the system does not depend on a commercial API that may change its pricing, terms, or availability. Second, it enables auditability: open weight models can be inspected, and a system running on institutional hardware can be held to institutional standards

of accountability. Third, it aligns with the European regulatory environment, from GDPR to the EU AI Act, in ways that cloud dependent architectures cannot.

This contribution resonates strongly with the keynote sessions. Goudey articulated the distinction between the model of technology development that optimises for friction removal and spectacular transformation and what he called the European specification: optimise for cognitive sovereignty, rigorous privacy, and sustainable integration [8]. Prégent's ALTAI framework, with its seven requirements for trustworthy AI [15], maps directly onto our design choices: privacy and data governance through student controlled profiles, transparency through the legible profile interface, non discrimination through reasoning style expression options, and human agency and oversight through the interaction model that prioritises reflection. Guerry's analysis of open source licensing and the meaning of sovereignty for AI systems [10] provides the legal and philosophical foundation for the infrastructure choice, while Beslin's framework [1] reminds us that the Social and Emotional dimensions (i.e. the student who feels seen rather than surveilled, the professor whose pedagogical insight is respected rather than replaced) are where the project will ultimately succeed or fail.

Any proposal that aspires to actionability must reckon with the conditions of its own implementation. We identify five conditions that would determine whether a system of this kind could operate in practice. The first is institutional commitment to sovereign infrastructure: a university that deploys this system must be willing to host open source models on its own servers or a shared institutional cloud, which requires technical capacity, financial resources, and a governance framework for model selection and updates. The UTC infrastructure already operational provides a proof of concept, but scaling to other institutions would require either a shared interuniversity infrastructure or technical capacity that not all institutions currently possess. The ILaaS project, which offers LLM inference and audio transcription services on demand for public higher education institutions using trusted servers hosted in labelled data centres and universities, represents a viable path forward for collective infrastructure investment.

The second condition is pedagogical integration. A learning companion that prioritises reflection cannot simply be layered onto existing curricula; it requires that educators design learning activities that assume its presence, that is, that leverage its profile driven, metacognitive capabilities rather than treating it as an add on. This implies professional development for faculty, codesign processes involving students and teachers, and a willingness to reconsider assessment practices that assume the absence of AI. The third condition is transparent governance. The profile raises questions that are as much political as technical: who decides what dimensions are tracked, how inferences are validated, what data is retained and for how long, and who has access to aggregated pedagogical intelligence. We propose a governance model that includes students, faculty, and institutional ethics bodies, with clear protocols for data access, profile correction, and the boundaries between personalisation and surveillance.

The fourth condition is disciplined evaluation. The success of this system cannot be measured by adoption rates or student satisfaction scores alone; it must be evaluated against longitudinal learning outcomes: retention at delayed intervals, transfer of knowledge to novel problems, and the development of self regulation skills. For this, a research design that compares cohorts using the system against matched controls should be tested, with hypotheses registered in advance and independent analysis. The fifth condition is interoperability. A learning profile that stays within a single institution is useful; one that travels with the student across institutions and platforms would be transformative. We advocate for the development of an open standard for student learning profiles, analogous to the Learning Tools Interoperability standard in the LMS ecosystem, that would

enable portability while preserving student control. Nipun Ranchhod Navadia's commentary on our proposal emphasised that the notion of a sovereign, student owned learning profile is strong and timely, especially with its GDPR framing, and that it introduces an interesting rebalancing of power between institution and learner [24]. We agree, and we note that the portability of the profile is what gives this rebalancing its force.

V. What Was Built: A Description of the Minimum Viable Product

The principles articulated above would remain speculative without a working system against which to test them. The remainder of this paper describes the minimum viable product we developed over the course of this challenge: a web-based learning companion that operationalises the metacognitive, sovereign, and Socratic commitments described in Section 4. The system is not a proof of concept in the narrow sense of a demonstration that certain functions are technically possible; it is a functional prototype with real users, real course materials, and real pedagogical logic embedded in its architecture. What follows is an account of what it does and how, written with enough specificity to make the design choices legible and evaluable.

The foundational architectural decision is what we call the Thinking Layer: a processing stage that sits between the student's input and the model's response, ensuring that every interaction is governed by pedagogical rules rather than by the model's default disposition to be helpful. This layer is not a filter that censors responses but a prompt architecture that conditions the model's behaviour through a layered system prompt: absolute identity and pedagogical rules at the base, a dynamic student state classifier in the middle, and the learner's profile and relevant course materials as the outermost context. The state classifier distinguishes three modes that the system detects continuously during a conversation. In the default mode, the system guides students through inquiry, asking questions that progressively narrow the gap between what the student knows and what the concept requires them to understand. When the system detects that a student is seeking a direct answer without engagement (through lexical patterns, question structure, and conversation history), it pivots into a redirection mode that acknowledges the request explicitly and proposes a structured path toward the same destination through reasoning. When persistent difficulty is detected, the system shifts into a scaffolding mode that provides more structured support, breaking the problem into smaller steps and making the intermediate logic visible. These three states are not experienced by the student as modes or labels but as variations in the system's conversational posture, calibrated to support rather than obstruct.

The second major component is the Student Learning Profile, a structured representation of each learner's cognitive characteristics that is updated incrementally throughout every session and synthesised comprehensively at its end. The profile tracks seven dimensions: dominant reasoning style, recurring misconceptions, demonstrated competencies, preferred exercise format, proximal zone indicators, engagement signals, and recent progress. These dimensions are not inferred from a single interaction but built from the accumulated evidence of the student's questions, explanations, errors, and corrections across sessions. The profile is stored in the student's account and made fully visible through a dedicated dashboard interface, where students can read the system's inferences about them, annotate entries they consider inaccurate. This design addresses directly the concern raised by Chinmay Das [23] about profiles that silently reinforce incorrect assumptions: the profile is not a background variable that the student never sees but a legible, editable document that functions as a cognitive mirror. Instructors, for their part, access only anonymised aggregate data across their cohort: the distribution of misconceptions, the topics generating the highest difficulty, and the engagement patterns that signal when a

given pedagogical approach is or is not working. No individual student's profile is accessible to the instructor without the student's explicit consent.

Course materials enter the system through an agentic retrieval-augmented generation architecture built on a vector database that enables search across the full corpus of uploaded documents. When a student asks a question, the system retrieves the most relevant passages from the course materials, queries again the database multiple time if needed (agentic RAG), injects them into the model's context as grounded truth, and constrains the system's response to what those materials support. It means the system's answers are bounded by the intellectual content of the course rather than by the model's general world knowledge, which prevents the well-documented drift toward plausible-sounding but curriculum-irrelevant responses. Students may also build personal document libraries by uploading their own notes and reading materials, which are indexed and made searchable alongside course materials for any course. The retrieval layer uses this combined corpus to personalise which passages are surfaced during a given session based on the student's profile and recent activity.

Exercise generation is handled by a dedicated agent that draws on both the course materials and the student's current profile to produce practice problems calibrated to the student's demonstrated level. The system supports two exercise formats. Multiple choice questions are graded immediately, with the system providing not a binary correct or incorrect verdict but a brief diagnostic explanation that targets the reasoning behind the wrong choice. Free response exercises, in which the student is asked to explain a concept, solve a multi step problem, or apply a procedure to a novel case, are evaluated by a specialised evaluation agent that reads each step of the student's response independently and returns targeted hints rather than a global score. A student who has correctly set up a problem but made an error in the third step receives feedback on that step, not a blanket assessment of failure; a student whose free text response demonstrates a correct result through flawed reasoning receives feedback on the reasoning rather than on the result. This step level feedback architecture operationalises the productive failure framework in a practical context: the student's incorrect attempt is treated as pedagogically informative rather than as an outcome to be corrected and discarded.

Study planning is available as a student-initiated feature. When requested, the system queries the course materials for topics relevant to the student's stated goal, drafts a structured progression from foundational to advanced concepts, and persists this plan to the student's dashboard, where each item can be expanded into a full exercise session. The plan is not a static schedule but a dynamic document: as the student completes items and the profile is updated, the system can identify which planned topics have been partially mastered through incidental discussion and flag them accordingly, adjusting the recommended focus. Professors can generate anonymised learning analytics reports at any point during a course, synthesising misconception patterns, engagement metrics, and coverage gaps across their cohort. These reports are generated without any access to individual conversation content, only to the profile-level inferences that have already been through the system's abstraction process, and they are designed to support pedagogical decision making rather than student evaluation.

The infrastructure on which all of this runs is deliberately institutional. The system is built on a FastAPI backend with a custom multi-agent orchestration layer, a React frontend, and a Supabase PostgreSQL database for user management and session history. The inference layer is designed around a fallback client architecture that routes requests across multiple model providers, including the local inference infrastructure operated by the Université de Technologie de Compiègne, and gracefully degrades when any single provider is unavailable. All Socratic and pedagogical functionality has been developed and validated on open-weight models, primarily from the Gemma and Mistral families, running on institutional hardware. The system is demonstrably functional at

this model tier; the performance gap relative to frontier proprietary models exists but, in the specific domain of structured Socratic tutoring with strong context conditioning, substantially smaller than general benchmarks would suggest, a finding consistent with the results reported by Favero et al. [6] on fine-tuned smaller models for Socratic dialogue. Student data does not leave the institutional perimeter, and the system is built for full GDPR compliance, with data export in structured format.

What this system is not is as important as what it is. It is not a replacement for the professor, whose role in designing learning sequences, maintaining intellectual authority over the course content, and building the social and emotional fabric of a learning community the system cannot replicate. It is not a cheating prevention tool, and its value does not depend on students being unable to circumvent it, for students who choose to go elsewhere for direct answers can do so freely (or use the direct mode); the system's purpose is to make the thinking path genuinely attractive, not to make the shortcut path inaccessible. And it is not finished. The tensions identified in Section 3, around equity in reasoning expression, around the boundary between helpful step by step understanding and paternalistic control, around the long-term validity of profile inferences, are design parameters that require longitudinal evaluation with real student cohorts before they can be considered resolved. What the prototype demonstrates is that a Socratic, sovereign, metacognitively aware learning companion is technically feasible on institutional infrastructure, that its pedagogical logic can be embedded in a working system rather than described only in theoretical terms, and that the design choices required to operationalise it are consistent with the European regulatory and ethical frameworks.

Conclusion

The question that frames this challenge, *what does learning mean in the age of AI?*, cannot be answered by technology alone, but neither can it be answered without technology. The AI that gives answers without building understanding is not neutral; it is a cognitive environment that shapes the minds that grow within it, just as surely as the architecture of a school shapes the learning that takes place there. The choice facing higher education is not whether to engage with AI but how: whether to accept the default interaction model of the commercial chatbot, optimised for engagement, retention, and data extraction, or to insist on a model that serves the purposes of education, which are the development of autonomous, reflective, self regulating thinkers.

Our contribution is a learning companion organised around a metacognitive student profile on sovereign infrastructure. It is one attempt to articulate that alternative. It is not the only possible one, and we do not claim to have resolved all the tensions that our analysis has surfaced. The relationship between cognitive friction and equity, the tension between sovereignty and capability, the legitimate concern about paternalism, and the open question of how to validate reasoning traces that may themselves be produced with the assistance of AI, these are not problems that a single prototype can solve. They are the terms of an ongoing debate that must involve educators, students, technologists, and policymakers, and that must be conducted with the same intellectual seriousness that we demand of our students.

What we offer is not a finished answer but an MVP with coherent direction: a set of principles grounded in cognitive science, tested through prototyping, and accountable to the regulatory and ethical frameworks that define European higher education. The path from principle to practice is long, and many questions remain open. But the cost of not pursuing it, of allowing the default commercial AI to become the de facto pedagogical infrastructure of higher education by inertia, is, we believe, higher than the cost of building something better.

Acknowledgments

The analysis presented in this report has drawn on the expert resources provided by the AI Grand Challenge organisation, including the meta-analytical work of Di Pietro and Castaño Muñoz on technology and less advantaged students, the white paper by Pr. Gahi on AI assistants in higher education, the critique by Arnaud Lévy of educational technology, the Référentiel général pour l'IA frugale published by the French Ministry of Ecological Transition [14], and the syntheses of existing studies on screen time and child development. The exchanges with the AI Grand Challenge community as in the comments of Mahdi Ayadi, Chinmay Das, Justine Cassell, and Nipun Ranchhod Navadia, as well as comments on other contributions, have shaped the arguments presented throughout. The keynote presentations by Alain Goudey, Alexandra Prégent, Bastien Guerry, and Caroline Beslin provided intellectual substance without which this proposal would be considerably poorer. We are grateful to the challenge organisers, the Project Advisory Group, and all participants whose questions and provocations pushed our thinking further than it would have gone alone.

References

[1] C. Beslin. Keynote presentation AI Grand Challenge 2026, Inria, 2026.

[2] R. A. Bjork and E. L. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pages 56–64. Worth Publishers, 2011.

[3] M. T. H. Chi, M. Bassok, M. W. Lewis, P. Reimann, and R. Glaser. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 1989.

[4] E. L. Deci and R. M. Ryan. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 2000.

[5] G. Di Pietro and J. Castaño Muñoz. A meta-analysis on the effect of technology on the achievement of less advantaged students. *Computers & Education*, 2025.

[6] L. Favero et al. Enhancing Critical Thinking in Education by means of a Socratic Chatbot In *ECAI 2024*, 2024.

[7] Y. Gahi. *L'Enseignement supérieur à l'ère des assistants IA*. Livre blanc, 2025.

[8] A. Goudey. Keynote presentation specification. AI Grand Challenge 2026, Inria, 2026.

[9] A. Goudey, R. Meissonier, et al. The legitimacy of generative AI in French higher education. *Communications of the Association for Information Systems*, 2026.

[10] B. Guerry. Keynote presentation AI Grand Challenge 2026, Inria, 2026.

[11] M. Kapur. Productive failure. *Cognition and Instruction*, 2008.

[12] J. A. Kulik and J. D. Fletcher. Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research*, 2016.

[13] A. Lévy. *L'Échec du numérique éducatif*. 2025.

[14] Ministère de la Transition Écologique. *Référentiel général pour l'IA frugale*. 2025.

[15] A. Prégent. Keynote presentation AI Grand Challenge 2026, Inria, 2026.

[16] L. Puech et al. Towards the Pedagogical Steering of Large Language Models for Tutoring: A Case Study with Modeling Productive Failure. In *Proceedings of ACL 2025*, 2025.

[17] H. L. Roediger and J. D. Karpicke. The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 2006.

[18] M. Stadler, M. Bannert, and M. Sailer. : Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 2024.

[19] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 2011.

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

[21] B. J. Zimmerman. Becoming a self-regulated learner: An overview. *Theory into Practice*, 2002.

[22] M. Ayadi. Comment on UTC EduTech proposal. AI Grand Challenge 2026, 29 April 2026.

[23] C. Das. Comment on UTC EduTech proposal. AI Grand Challenge 2026, 21 April 2026.

[24] N. R. Navadia. Comment on UTC EduTech proposal. AI Grand Challenge 2026, 10 May 2026.

[25] J. Cassell. Comment on UTC EduTech proposal. AI Grand Challenge 2026, 30 April 2026.