

A Strategy for AI & Education

A blueprint for educational institutions to approach artificial intelligence strategically, in collaboration with **Empathy.AI**.



Prepared by	Empathy.AI
Intended audience	Educational leadership, pedagogical and technology teams
Published	September 2026
Nature of this document	Strategic guidance, not a commercial offer

This document is a blueprint, not a prescription. It is intended to help educators, school leaders and technology teams think through how to approach artificial intelligence in education deliberately, and to show, concretely, what a governed, privacy-first, locally owned approach looks like in collaboration with Empathy.AI. Empirical claims throughout are referenced to the published literature; full references appear on pages 18 and 19.

Contents

The Cognitive Challenge, and Empathy.AI's Response	3
What the Research Shows: Cognitive Debt and the Case for Governance	4
From Evidence to Design	5
Three Reframes: What Governed AI Makes Possible	6
Empathy.AI for Education	7
Four AI Extensions of AI in Education	8
Empathy.AI Spaces: Software Ready to Use	9
A Capacity Architecture Already Proven	10
An Engagement Timeline	11
Case Studies: Empathy.AI Spaces in Production	12
Educational Data of Minors: Governance by Design	13
A Private AI Utility, Not a Single Chat Session	14
Relevant Economics for Institutional Teams	15
Frequently Asked Questions	16
References	18
A Blueprint, Not a Prescription	20
Annex A · The Local AI Workstation: Architecture and Possibilities	21
Annex B · Empathy.AI Spaces: Advantages	23
Annex C · Case Study: Putney High School	24

The Cognitive Challenge, and Empathy.AI's Response

The Challenge: Artificial Intelligence and Cognitive Development

We are told that the future of education is already here. Machines will teach, algorithms will grade, and AI will make every lesson more efficient. What a dull future that promises, if we are not careful.

The real danger is not that machines will optimise teaching or replace teachers. It is that AI will entrench the very system education has long been trying to outgrow: a system of conformity, where learning is reduced to metrics and students to measurable, predictable things (Selwyn, 2019; Holmes & Tuomi, 2022). In such a system children are trained to produce efficiently rather than think critically, and learning stops being a journey of discovery and becomes a race to produce answers. That is obedience, not education.

Big Tech's AI in education is built on a foundation of control: trained on data no student or teacher has consented to, and sustained by constant data tracking and commercial subscriptions (Zuboff, 2019; Williamson, 2017). What gets made efficient is attention, engagement, and the platform's own profit, not learning (Bender et al., 2021).

The Response: Empowerment, Control, Agency, and Autonomy

Empathy.AI offers a different path: AI as a tool, designed to function in line with human agency, governance and control. It can run locally, on an institution's own servers, on its own campus. No data needs to leave the institution. No student needs to be tracked. It can be grounded in what the institution chooses: its own textbooks, lesson plans, curricula, and values.

The goal is to return judgement to the people who should exercise it. The teacher becomes a guide and storyteller again rather than the administrator of a system they do not control, and the student relates to knowledge on their own terms rather than being measured and predicted by it. That is why the question this blueprint is built around is not “how can AI improve education?” but “how can we better relate with knowledge?”.

The technical approach set out in this blueprint, a locally owned AI engine running Empathy.AI Spaces, is the concrete expression of this posture: empowerment instead of dependency, control instead of opacity, agency instead of silent automation, institutional autonomy instead of surveillance.

Source: AI in Education: An Empathetic Vision for Creativity, Art, and Liberation, Empathy.AI

What the Research Shows: Cognitive Debt and the Case for Governance

In 2025, researchers at the MIT Media Lab published one of the first controlled studies measuring what happens inside a learner's brain when AI assistance is introduced into writing (Kosmyna et al., 2025). Over four months, 54 participants wrote essays using a large language model, a search engine, or no tool at all, while researchers recorded brain activity via 32-channel EEG and tracked memory recall, linguistic originality, and each participant's sense of ownership over their work.

The finding was unambiguous on one point: **brain connectivity scaled down with the amount of external support.** The unaided group showed the strongest, most widely distributed neural networks; the search-engine group showed intermediate engagement; the LLM group showed the weakest coupling of all. Reported ownership of the finished essay followed the same gradient. The direction of the result is consistent with the wider literature on cognitive offloading (Risko & Gilbert, 2016; Gerlich, 2025).

The researchers' own term

"This pattern reflects the accumulation of **cognitive debt**, a condition in which repeated reliance on external systems like LLMs replaces the effortful cognitive processes required for independent thinking. Cognitive debt defers mental effort in the short term but results in long-term costs, such as diminished critical inquiry, increased vulnerability to manipulation, decreased creativity." (Kosmyna et al., 2025)

78%

of habitual LLM users could not quote a single line from an essay they had personally written minutes earlier, once AI support was removed (Kosmyna et al., 2025)

11%

was the equivalent failure rate for participants who had written unaided from the start, a sevenfold gap in retained ownership of their own work

The study also reports a finding that speaks directly to how AI should be designed for education: timing matters as much as exposure. In a fourth session, participants who had written unaided for three sessions were then given AI access for the first time, and showed a network-wide spike in neural connectivity stronger than any session in the habitual-LLM group. Sequence was the variable that moved the result.

The study's own design implication

"These findings support an educational model that delays AI integration until learners have engaged in sufficient self-driven cognitive effort. Such an approach may promote both immediate tool efficacy and lasting cognitive autonomy." (Kosmyna et al., 2025)

Independent causal evidence points the same way. In a field experiment with roughly one thousand secondary mathematics students, an unguarded GPT-4 interface lifted practice performance by 48%, yet once access was removed those students performed 17% worse than peers who had never used it; the same model behind teacher-designed guardrails lifted performance by 127% and produced no subsequent harm (Bastani et al., 2025). The design of the interaction, not the model, determined whether learning survived the tool's removal.

From Evidence to Design

The MIT study also names why today's dominant AI tools default to the wrong sequence: the LLM “undeniably reduced the friction involved in answering participants' questions... this convenience came at a cognitive cost, diminishing users' inclination to critically evaluate the LLM's output... What is ranked as 'top' is ultimately influenced by the priorities of the LLM's shareholders” (Kosmyna et al., 2025). A system optimised for engagement has no incentive to withhold itself until a learner has struggled first; friction is exactly what it is built to remove (Zuboff, 2019).

A commercial AI product has no mechanism to enforce the sequence the research points to, and no incentive to build one: convenience is its value proposition. A governed, locally owned engine can, because governance and sequencing are exactly what its controls are for.

What the research recommends	What governed local AI provides
Delay AI assistance until self-driven effort has occurred (Kosmyna et al., 2025; Bastani et al., 2025)	Agentic Governance: educators configure when and how AI activates in a learning pathway, rather than an always-on assistant with no gate
Counter shallow encoding from consuming answers passively (Slamecka & Graf, 1978; Karpicke & Blunt, 2011)	Provenance & Sources: every answer declares its reasoning and traces to authors and curators, so a student engages with where knowledge came from, not just the output
Remove engagement-optimised, shareholder-ranked defaults (Zuboff, 2019; Bender et al., 2021)	No surveillance economics: a locally owned model has no engagement metric to maximise and nothing to rank by commercial priority
Preserve learners' sense of ownership over their own work (Ryan & Deci, 2000)	No anthropomorphism, explicit boundaries: the system is positioned as a tool a student directs, not an authority that answers for them

Local AI, on its own, does not resolve cognitive debt. The same risk exists on infrastructure an institution owns if the tool is left ungoverned. What ownership makes possible is the missing piece: the ability for an institution to actually design and enforce the sequence the evidence calls for, instead of inheriting a vendor's default.

Three Reframes: What Governed AI Makes Possible

Cognitive debt is not an inherent property of artificial intelligence, and neither is its opposite. Both are properties of design: of whether a system is built to fade, to preserve the right kind of difficulty, and to invite scrutiny instead of discouraging it. Read against the evidence above, governance reframes AI's relationship to learning along three specific axes.

01 Cognitive Scaffolding, not Offloading

instead of outsourcing thinking to AI, AI as a temporary structure withdrawn as competence grows

Offloading hands the whole cognitive task to a system and collects the output; the effortful thinking that would have built understanding simply never happens, the way a calculator outsources arithmetic a student was supposed to be learning (Risko & Gilbert, 2016). Scaffolding, a concept older than AI by decades, holds a task steady long enough for a learner to reach a skill level slightly beyond what they could manage alone, within Vygotsky's zone of proximal development, and is deliberately withdrawn as competence grows (Vygotsky, 1978; Wood, Bruner & Ross, 1976). Same tool, opposite cognitive outcome; the determining factor is whether the system is designed to fade.

Enabled by: Agentic Governance. Educators configure when AI assists and what kind of help it is allowed to give, rather than inheriting an always-on assistant with no gate.

02 Neuroplastic Enhancement, not Intellectual Atrophy

instead of mental skills wasting away, new neural pathways form by tackling harder problems once busywork is removed

Cognitive load theory distinguishes *germane* load, the productive struggle of connecting new information to what a learner already knows, from *extraneous* load, the friction of formatting, retrieval and restating what is already understood (Sweller, 1988; Sweller, van Merriënboer & Paas, 1998). What atrophied in the MIT study were germane-load functions: memory encoding, source recall, the ability to reconstruct one's own argument. The cause was not AI reducing effort; it was AI reducing the *wrong* effort. Remove only the extraneous load, and a learner tackling harder, more ambiguous problems forms new pathways rather than fewer of them, which is also what the literatures on the generation effect, retrieval practice and desirable difficulties predict (Slamecka & Graf, 1978; Karpicke & Blunt, 2011; Bjork, 1994).

Enabled by: Provenance & Sources. Declared reasoning on every answer preserves the germane load of evaluation even as it strips out the extraneous load of retrieval.

03 Active Metacognition, not Passive Consumption

instead of blindly accepting AI answers, users learn to question, prompt and audit outputs, thinking about their own thinking

The MIT researchers also observed that LLM users showed a “diminished inclination to critically evaluate the LLM's output.” This connects directly to why the other two effects compound: an AI never interrogated is an AI fully depended on, whether or not that was the intent. The countermeasure is metacognition, asking where an answer came from, what is missing, what a different framing would look like; and metacognition is a habit built by environments that require it repeatedly, not a trait learners either have or lack (Flavell, 1979; Zimmerman, 2002).

Enabled by: transparent, cited answers. A black box trains a learner not to ask questions, because there is nothing to ask; a contestable, sourced answer makes contesting it the obvious next move.

Commercial AI, optimised for engagement rather than a learner's development, has structural reasons to land on the wrong side of all three reframes: offloading is stickier than scaffolding, and passive consumption keeps people coming back in a way active interrogation does not (Zuboff, 2019). Governance is what lets an institution choose otherwise.

Empathy.AI for Education

The premise of Empathy.AI's technology is that intelligence lives in the knowledge of each context, and the model is only the tool. In education, this means building AI systems an institution can see, audit, and govern: systems that students and staff can actually trust.

“Building self-hosted AI you can trust.”
Source: Edge AI for Education, Empathy.AI

The Problem: Commercial AI in the Classroom

Commercial AI in the classroom rests on models trained on data that students and teachers do not control, and requires constant connectivity and commercial subscriptions. Its design is organised around engagement metrics that may not align with learning outcomes (Zuboff, 2019; Kasneci et al., 2023), and it is sustained by data collection that raises questions of privacy and consent for minors (Livingstone, Stoilova & Nandagiri, 2019). The result is dependency on external platforms and vendors, and an erosion of institutional autonomy (Williamson, 2017).

Empathy.AI's Values for Education

Student Data Sovereignty

Local execution on the institution's own infrastructure, with no external transmission. All personal data processing occurs exclusively within the EU, in line with the GDPR (Regulation (EU) 2016/679).

Verifiable Transparency

Every answer shows the exact source that supports it, essential for teaching research skills, and the system is able to say “I don't know” instead of hallucinating.

Explicit Ethical Boundaries

Questions the system should never answer are explicitly defined; the final decision always remains with teachers and students.

Independence from Commercial Platforms

No surveillance economics: the system reflects the institution's own pedagogical values and educational standards, not a technology vendor's objectives.

Open and Auditable Models

Inspectable weights and a model version that stays fixed until deliberately changed. Where deep provenance is required, options such as OLMo 3.1 32B offer fully open training data (Groeneveld et al., 2024).

The Local AI Workstation as an Expression of This Vision

Permanent, owned terms: data stays inside the institutional boundary, the model version is stable, price is fixed at purchase, weights are inspectable, and continuity is independent of any vendor's roadmap.

Four AI Extensions of AI in Education

A well-designed AI engine for education turns learning evidence, activities, assessments, self-evaluations and teacher observations, into actionable insight. In practice, this typically spans four capability lines:

Agentic Governance
Drive how learning pathways are designed, associating challenges and activities in line with an educator's configurations and controls.

Learning Pattern Analysis
Analysis of patterns and competency evolution from the evidence generated throughout the learning process.

Source Declaration
Declaration of sources accompanying the reasons that explain answers, tracing to authors and curators.

Language Technologies and Generative AI
Automatic generation of progress reports, pedagogical recommendations and personalised feedback.

What This Requires: A Structured Collaboration

Turning these four extensions into a working, governed engine typically requires specialised expertise in software architecture and AI integration, centred on three deliverables: designing the technical architecture of the engine, defining its integration mechanisms with the institution's existing learning platform, and providing technical support during implementation of the AI components.

Aligning With an Institution's Own Objectives

Common educational objective	Empathy.AI's response
Turning learning evidence into actionable information	A private knowledge space that indexes that evidence with local vector retrieval
Personalised recommendation of pathways, challenges and resources	A suggestion service built on each student's competency profile, with no usage meter
Analysis of learning patterns and progress	Deep multi-document research (32k–64k token contexts) over the evidence history
Predictive models (support needs, stagnation, participation)	Background agents and authorised overnight windows for recurring detection and anticipation
Generative AI: reports, recommendations, feedback	Local inference (Qwen3.8-27B or OLMo 3.1 32B) with mandatory source disclosure in every report
Institutional autonomy and protection of minors' data	On-premises deployment, restricted egress by default, EU GDPR compliance

In short: an institution wants AI to work for learning; Empathy.AI builds infrastructure where AI works with people, on their own terms.

Empathy.AI Spaces: Software Ready to Use

A Local AI Workstation provides compute capacity. The real value an institution adopts is **Empathy.AI Spaces**: a private knowledge application, already built and operating, on top of which an AI engine can be deployed as an instance tailored to the institution's own educational domain.

Knowledge

An Institution's Own Source Library

Evidence, rubrics, curricula and pedagogical documentation are processed as the Space's Knowledge for retrieval.

Chat

Grounded, Cited Answers

Every answer traces back to the documents that produced it: a verifiable partner in analysis, not a black box.

Controls

Governance Without Prompt Engineering

Always-on platform guardrails, plus the institution's own rules on tone, format and domain context, which staff can enable, review or disable.

Adaptive Mode

Real Agentic Capability

For broader questions, it runs several retrieval steps before answering, and can generate downloadable documents, reports and guides, from the Space's own knowledge.

History & System

Traceable and Transparent

Manageable, exportable conversations; a read-only view of how the engine is configured: models, pipeline, ingestion.

Privacy & Sovereignty

Inference Within the Boundary

Every step runs inside the Space. Conversations are never used to train any model, whether the institution's own or a third party's.

*An AI-in-education engine does not need to start from zero: it can be built **on Empathy.AI Spaces**, the same platform already governing Knowledge, Chat, Controls, Adaptive and Sources in live deployments. The work required is adapting that platform, and its four capability lines, to an institution's own competency domain.*

Source: About Empathy.AI Spaces, Empathy.AI

A Capacity Architecture Already Proven

The four capability lines above map naturally onto four capacity lanes that already form part of the Local AI Workstation's reference design. There is no need to start from zero: a shared-engine architecture of frequent retrieval, occasional deep analysis, continuous monitoring agents and batch generation can be adapted to any institution's educational domain. Retrieval quality rests on contextual enrichment of document chunks before indexing, shown to reduce retrieval failure rates by up to 67% against standard chunking when combined with reranking (Anthropic, 2024).

Educational capability line	Engine capacity lane	Typical behaviour
Recommendation and personalisation	Knowledge-work RAG	Frequent, mid-context queries: competency profile plus available resources
Pattern and progress analysis	Deep research	Broad context, occasional sessions: historical evidence per student or cohort
Predictive models	Background agents	Scheduled monitoring: stagnation or low-participation alerts
Generative AI (reports, feedback)	Overnight / batch processing	Batch generation of progress reports and personalised feedback

Reference Architecture

Your Learning Platform

activities · assessments · self-evaluations · teacher observations

Local AI Workstation — Sovereign, On-Premises Environment

01 Private knowledge space: document ingestion, PostgreSQL/pgvector, self-hosted vLLM

02 Local retrieval: Qwen3-Embedding-4B, reranking, permission control

03 Interchangeable reasoning engines: Qwen3.8-27B · OLMo 3.1 32B, selectable per space or task

04 Educational AI workload: recommendation · patterns · prediction · reports · feedback

NVIDIA RTX PRO 4500 Blackwell · 32GB VRAM · 896 GB/s · ~300 W

Output to staff: reports and feedback with traceable sources · every recommendation cites the evidence retrieved · nothing leaves the institution

Capacity bands are planning estimates, not throughput guarantees. Exact model quantisation, KV-cache precision, serving engine, context length and latency targets should be benchmarked before deployment.

An Engagement Timeline

A structured collaboration with Empathy.AI typically unfolds across three sequential phases. The timeline below is indicative; durations can compress or extend depending on an institution's existing platform, data readiness and internal capacity.

01

Research & Architecture

Months 1–3. Technical design of the AI engine: models, capacity lanes, evidence storage and governance boundaries, grounded in the institution's actual data and workflows.

02

Integration Design & Prototype

Months 4–7. Definition of integration mechanisms with the institution's existing learning platform: API contracts, evidence formats, authentication and data flows, validated through a working prototype.

03

Implementation Support

Months 8–12+. Ongoing technical support during rollout of the AI components, with joint validation alongside the institution's own technology team. Extendable in further tranches as needs evolve.

Typically outside this scope

Development of the institution's core learning platform itself; third-party software licences outside the defined open stack; hardware acquisition, unless separately scoped as its own workstream.

Case Studies: Empathy.AI Spaces in Production

Empathy.AI Spaces already operates in production, combining knowledge retrieval with agentic capability, including the autonomous generation of documents, the same pattern an AI-in-education engine needs for progress reports and pedagogical recommendations.

Molly Vs The Machines

mollyvsthemachines.com

Knowledge Retrieval

Agentic Capability

Document Generation

The official website of the documentary on child safety and Big Tech platforms. It deploys an Empathy.AI Space combining **Knowledge Retrieval** over the film's transcript and reference sources on surveillance capitalism, with an **agentic** capability: the "Create Screening Guide" function uses Adaptive mode to generate, on demand, a downloadable educational guide built from the Space's own knowledge.

Relevance to education: demonstrates, in production, exactly the pattern an educational AI engine needs to replicate: evidence-grounded answers, plus autonomous generation of a final pedagogical document, equivalent to the progress reports and personalised feedback described in the fourth pillar above.

Project Gutenberg AI

projectgutenberg.empathy.ai

Large-Scale Knowledge Base

Semantic Search

Private Infrastructure

A semantic discovery system developed by Empathy.AI in collaboration with Project Gutenberg, the world's oldest public-domain digital library. It allows exploration of more than 78,000 works through conversation grounded in the actual text, by theme, emotion or narrative structure, not only keyword, built on the same Knowledge Engine (agentic RAG) that underpins Empathy.AI Spaces, and run entirely on private infrastructure.

Relevance to education: validates the platform's ability to operate over a large, heterogeneous knowledge base without losing precision or governance, the same challenge posed by a large-scale competency evidence base, while keeping privacy by design in every query.

Putney High School

Girls' Day School Trust (GDST) · London

On-Campus Deployment

Age-Staged Access

Governed by School Policy

A selective all-girls school in the GDST family runs Empathy.AI Spaces on a dedicated Local AI Workstation physically installed on campus, filling a gap the school had identified in its own published curriculum policy: no named, school-authored AI programme. Access is age-staged: guided, safe first contact at Year 9 inside a closed, source-cited perimeter, and programmatic, student-built use at Year 13, with pupils' own code running in an isolated sandbox, all under identity and retention policy the school itself writes.

Relevance to education: demonstrates the age-staged governance model most directly transferable to any school setting, from closed first contact to supervised student-led development, on infrastructure where nothing leaves the building. *The full case study is presented at length in Annex C.*

Educational Data of Minors: Governance by Design

The learning evidence and competency profiles that feed an educational AI engine mostly belong to minors. In that context, keeping processing within an institution's own boundary stops being a preference and becomes a design condition demanded by data protection law: the GDPR's requirements of data protection by design and by default apply in full to the profiling of pupils (Regulation (EU) 2016/679), the UK's Age Appropriate Design Code requires high-privacy defaults for services likely to be accessed by children (ICO, 2020; Livingstone, Stoilova & Nandagiri, 2019), the EU AI Act treats education as a high-risk domain (Regulation (EU) 2024/1689), and UNESCO's global guidance calls for mandated data-privacy protection and institutional validation of AI tools before classroom use (Miao & Holmes, 2023).

Governance in Action

Who Holds the Terms

An AI engine deployed on a Local AI Workstation, compared with an equivalent engine built on a cloud AI API.

Dimension of control	Workstation (local)	Cloud API (any provider)
Data stays inside your boundary nothing sent to a third-party API	✓	✗
Model version fixed until you change it no silent swaps, deprecations or retraining	✓	✗
Pricing fixed once you own it a vendor can reprice a live endpoint at will	✓	✗
Weights fully auditable you can inspect what is actually running	✓	✗
Continuity independent of a vendor's roadmap no forced migration when a model is sunset	✓	✗
Governed by policy you wrote not a Terms of Service you clicked through	✓	✗

The real question is not “Can AI kill us?”

Public debate frames AI danger as an event: some kind of catastrophe where the system breaks containment and acts against humanity. That framing is misleading. The more plausible harm is gradual: innovations that erode creativity, judgement and autonomy by narrowing free will, delivered by default and without deliberation (Bender et al., 2021; Gerlich, 2025).

A locally owned AI engine is one concrete defence against that erosion: the institution, not a platform, decides what the tool is for.

A Private AI Utility, Not a Single Chat Session

Built for staff. One knowledge engine. Continuous retrieval, research and agentic work, without a per-question meter.

“Cloud bills every act of reading. A local engine rewards keeping the machine busy.”

The economics begin with sustained knowledge work. A Local AI Workstation is not intended to win against casual public-chat usage; it is intended to replace recurring private AI capacity for institutions whose knowledge, workflow and governance needs justify an owned execution layer. Below that threshold, low-cost public APIs may remain the simpler economic choice.

Assumption	Planning value
Acquisition cost	€10,000
Amortisation period	36 months
Active load	300 W
Electricity	€0.30/kWh
Team size	Shared capacity, built for staff teams
Local operating cost	€302–343/month

One Shared Queue, Many Jobs

A shared private-capacity engine, not a one-user chatbot. Workload classes are scheduled in one shared queue.

Capacity lane	Typical context	Planning concurrency	Purpose
Knowledge-work RAG	16k–32k tokens	3–5 active sessions	Daily evidence retrieval, drafting, analysis and synthesis
Deep research	32k–64k tokens	1–2 active sessions	Multi-document review, investigation, audit and long-form reasoning
Background agents	Variable	1–3 scheduled streams	Monitoring, extraction, enrichment, tool use and autonomous research
Overnight processing	Variable	Uses off-peak capacity	Document classification, synthesis, indexing and recurring analysis

Capacity
32GB private VRAM · 896 GB/s memory bandwidth · one shared queue, many jobs.

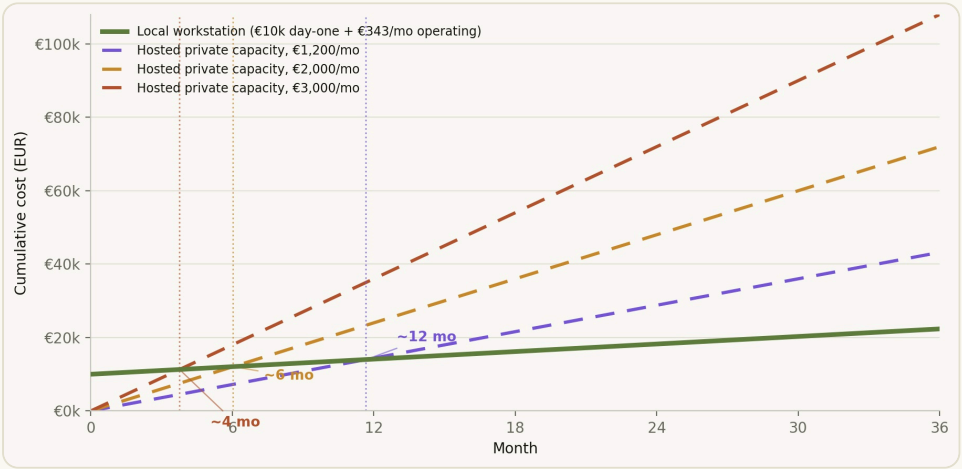
Relevant Economics for Institutional Teams

Comparator: equivalent private-capacity service, not the lowest publicly advertised, best-effort API rate.

Utilisation profile	Operating pattern	Monthly tokens	Cloud-equivalent spend
Knowledge-work team	A team; 20 questions/day; ~32k avg. context; recurring retrieval, selective automation	~290M	€350–900/mo
High-utilisation Space	A team; 20 q/day; ~32k context; 2–4 continuously active agents; regular batch work	~600M–1B	€1,000–2,500/mo
Research & automation hub	A team; 32k–64k deep-research tasks; 4–8 continuous agents; overnight processing	~1B–3B+	€2,000–5,000+/mo

Payback View

Equivalent hosted spend	Local cost at 24/7	Monthly saving	Simple payback on €10,000
€900/mo	€343/mo	€557/mo	~18 months
€1,200/mo	€343/mo	€857/mo	~12 months
€2,000/mo	€343/mo	€1,657/mo	~6 months
€3,000/mo	€343/mo	€2,657/mo	~4 months
€5,000/mo	€343/mo	€4,657/mo	~2 months



The local line begins with the €10,000 acquisition and rises with the full local technical operating cost (€343/month). Hosted lines assume no upfront cost and bill purely by the month, and can reprice, retire or change terms on the vendor's own schedule.

Frequently Asked Questions

Student and Institutional Data

01 What is a local AI workstation?

A dedicated computer optimised for ownership, offering a private, self-contained and sovereign way to relate with knowledge. It is designed to safely converse with an institution's knowledge and run agentic workflows without undermining its autonomy. Documents, prompts, retrieved evidence, embeddings, model outputs and the reasoning loop remain within the workstation; they are not sent to a public AI API for inference. What the AI becomes depends on the institution, not on what Big Tech decides.

02 Does data leave the institution?

No. Documents, prompts, retrieved evidence, embeddings, model outputs and the reasoning process remain within the workstation. It holds the knowledge, retrieves relevant evidence and performs reasoning locally. Privacy is not an optional setting or a promise made by an external provider; it is how the system operates.

03 Is this just a private version of ChatGPT?

No. Chat is only one way of meeting a system. The workstation creates a private Space where staff can search, retrieve, question, compare, write, investigate and run carefully governed agentic workflows around the institution's own knowledge, within limits it sets.

04 Why does data sovereignty matter?

Because control over knowledge shapes control over decisions. An institution's documents, questions, working drafts and internal reasoning do not become prompts processed by a public AI provider, subject to that provider's infrastructure, retention policies, commercial incentives or changing terms. The institution decides what goes in, who may use it, what is retained, and what its agents may do. This is also the posture data protection law expects for minors' data (Regulation (EU) 2016/679; ICO, 2020).

05 Can confidential records and internal documents be used?

Yes. The workstation can bring approved records, policies, technical documentation and repositories into a searchable Space, where they remain within the workstation. Retrieval, reasoning and answers all happen locally. Confidential knowledge is never sent to a public AI API simply because someone asks a question.

How Retrieval Works Inside a Private, Local Space

06 What does RAG mean in practical terms?

RAG means the workstation looks for relevant evidence before it answers. Documents are searched locally; the system selects the most relevant material, brings it into the active conversation and reasons over it locally. Sources, retrieval process and answer all remain within the workstation.

07 Why are large context windows useful?

They let the workstation hold more relevant evidence in view at once: comparing a student's progress across terms, synthesising multiple reports, or tracing a pattern across an archive without fragmenting the work into disconnected prompts.

08 Does an enormous knowledge base sit inside the GPU?

No. Knowledge lives in a structured, governed local archive alongside the workstation's operating environment. On a question, the system retrieves the most relevant material, tests and reranks it, then brings a carefully selected body of evidence into the model's working context.

09 How does it avoid being overwhelmed by a huge knowledge base?

It narrows attention computationally. A good retrieval system uses metadata, permissions, semantic search, keyword signals and reranking to find relevant evidence, offering the model a bounded, useful context rather than an indiscriminate flood of documents (Anthropic, 2024).

10 Can the system show its sources?

Yes. It is configured to reveal and return the exact documents, passages or records that participate in every response or action, making it a partner in inquiry rather than a black box.

Frequently Asked Questions

Concurrent Use by Staff

11 Can more than one person use it at once?

Yes. The workstation is designed as shared local capacity: a private resource that can support multiple staff, research workflows and scheduled agentic tasks simultaneously; the institution decides how it is shared.

12 How many people can use it in parallel?

That depends on the depth of the task. For everyday knowledge work, several people can work concurrently. More demanding research, with larger evidence packs and longer reasoning, is best treated as a prioritised or queued activity.

13 What happens if everyone asks a difficult question at once?

The system gives work a place in a shared queue. Short, everyday questions stay responsive, while longer research tasks are admitted deliberately and prioritised transparently.

14 Why does memory bandwidth matter?

It affects how fluidly the workstation moves between model knowledge, working context and concurrent requests, supporting a more responsive shared Space where everyday questions, deeper research and background work can coexist.

15 Can it work overnight?

Yes. When staff step away, it can continue approved work: processing new documents, updating indexes, preparing research, or running carefully bounded agents, while people retain full control over what is permitted.

When Ownership Makes Sense

16 Is local AI always cheaper than cloud AI?

No. For occasional, low-volume use, public cloud services may cost less on narrow token-price terms. A workstation is built for a different need: sustained, private, context-heavy knowledge work. Its value is not only a lower unit price, but the ability to use AI without handing an institution's knowledge and choices to an external platform.

17 What does “no per-question meter” really mean?

After acquiring the workstation, asking another question does not create another public-API token invoice. Real costs remain, hardware, electricity, maintenance, but every retrieval, research task and approved workflow is processed inside infrastructure the institution owns.

18 Why does RAG change the cloud-cost conversation?

Answering a serious question may require reading thousands of tokens of source material. In a cloud model, that reading is billed repeatedly; in a local Space, it uses capacity the institution controls. It changes the relationship from paying per act of reading to managing a shared resource.

19 When does a local AI workstation make commercial sense?

When the need goes beyond occasional chat and becomes sustained private knowledge work: daily use of confidential material, recurring retrieval and research, a need for source-grounded answers, and a desire to keep institutional knowledge inside infrastructure it governs.

20 What is an institution really buying beyond a GPU?

The conditions for a different relationship with AI: a private Space where models, documents, evidence, permissions and agentic actions are organised around human responsibility, useful without becoming a distant authority; powerful without becoming extractive.

References

All empirical and theoretical claims in this blueprint are referenced to the following published literature. In-text citations use author and year.

- Anthropic. (2024). *Introducing contextual retrieval*. Anthropic Engineering. <https://www.anthropic.com/engineering/contextual-retrieval>
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. <https://doi.org/10.1073/pnas.2422633122>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
- Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1), 6. <https://doi.org/10.3390/soc15010006>
- Girls' Day School Trust. (n.d.). *The Girls' Futures Report*. GDST. <https://www.gdst.net>
- Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., et al. (2024). OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 15789–15809). ACL. <https://doi.org/10.18653/v1/2024.acl-long.841>
- Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, 57(4), 542–570. <https://doi.org/10.1111/ejed.12533>
- Information Commissioner's Office. (2020). *Age appropriate design: A code of practice for online services*. ICO. <https://ico.org.uk>
- Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018), 772–775. <https://doi.org/10.1126/science.1199327>
- Kasnezi, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, 41(2), 75–86. https://doi.org/10.1207/s15326985ep4102_1
- Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task. *arXiv preprint arXiv:2506.08872*. <https://arxiv.org/abs/2506.08872>
- Livingstone, S., Stoilova, M., & Nandagiri, R. (2019). *Children's data and privacy online: Growing up in a digital age*. London School of Economics and Political Science.
- Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. <https://doi.org/10.54675/EWZM9535>

References

- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230–253. <https://doi.org/10.1518/001872097778543886>
- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 (General Data Protection Regulation). *Official Journal of the European Union*, L 119, 1–88.
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689.
- Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.
- Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6), 592–604. <https://doi.org/10.1037/0278-7393.4.6.592>
- Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645–3650). ACL. <https://doi.org/10.18653/v1/P19-1355>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/A:1022193728205>
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
- Williamson, B. (2017). *Big data in education: The digital future of learning, policy and practice*. Sage.
- Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2
- Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.

Documentary references (Empathy.AI)

Local AI Workstation, v3 (technical dossier: hardware, installed models, system architecture, economics, capacity and governance) · *Edge AI for Education* (workshop programme on sovereign AI for education: values, minors' privacy, transparency and ethical boundaries) · *AI in Education: An Empathetic Vision for Creativity, Art, and Liberation* · *AI Governance in the Perspective of Empathy AI* (July 2026) · *About Empathy.AI Spaces* (product overview and deployment options).

A Blueprint, Not a Prescription

This document is intended as a starting point, not a fixed script. Every institution's data, platforms and pedagogical priorities differ, and the right architecture, capacity, and pace of adoption should be shaped around them. What stays constant is the underlying posture: empowerment instead of dependency, transparency instead of a black box, and governance that an institution writes for itself rather than accepts by default.

Empathy.AI and Its Net Zero Headquarters

Gijón, Asturias, Spain

A bioclimatic building, powered by solar energy



Empathy.AI is a foundational AI platform within **Empathy Holdings**, built to develop ethical, sustainable and privacy-first AI solutions for commerce, legal, government, healthcare, finance and publishing. Its main infrastructure runs on **net-zero energy** in a bioclimatic building in Gijón, Asturias, powered by solar panels, ensuring that AI development does not come at the cost of the planet, a consideration the research community has urged on the field (Strubell, Ganesh & McCallum, 2019).

Net-Zero Operations

Solar-powered infrastructure, with no carbon footprint associated with compute.

Edge AI for Education

Workshops where educators visit the physical infrastructure in Gijón to see, touch, and understand how the AI works.

Aligned With the UN SDGs

Climate Action (SDG 13), Responsible Consumption (SDG 12), and Sustainable Cities (SDG 11).

Empathy.AI rejects the “compute race” model: its platform is designed to do more with less, not more with more. Sustainability is not a side project; it is a core value.

Starting the Conversation

This blueprint is designed to be worked through together. Empathy.AI can support an institution's leadership and technology teams in shaping a strategy grounded in the frameworks set out above.

[Begin a Strategic Conversation with Empathy.AI →](#)

The Local AI Workstation: Architecture and Possibilities

A complete technical reference for the local AI infrastructure that underpins the approach set out in this blueprint.

“It is a visual allegory, a tool for hope: a dedicated computer optimised for ownership, offering a private, self-contained and sovereign way to relate with knowledge, free from Big AI.” (Local AI Workstation, Empathy.AI)



Vertical GPU riser mount, internal build



The assembled workstation, finned intake face

01 · Compute

Dedicated GPU, on Your Desk

32GB professional Blackwell card. No shared tenancy, no public cloud.

02 · Models

Authored Knowledge Only

Relate with knowledge in line with your terms.

03 · Data

Privacy-First

Documents and prompts are processed locally. Nothing leaves the building.

04 · Spaces

Full Agentic Capability

Adaptive mode and agentic tool-use, running end-to-end, on-premises.

Hardware Specification

Component	Model	Notes
Motherboard	ASUS ROG STRIX X870-I Gaming WiFi	Mini-ITX
CPU	AMD Ryzen 7 9700X	
CPU cooler	Noctua NH-L12Sx77	Low-profile
Memory	Kingston FURY Beast DDR5 32GB (2×16GB)	
GPU	NVIDIA RTX PRO 4500 Blackwell (PNY)	32GB VRAM, 896 GB/s
Storage	Kingston NV3 1TB NVMe SSD	
PCIe riser	JMT PCIe 5.0 x16 180°	
Power supply	Cooler Master GX II Gold 850W	Modular
Chassis fans	2× Noctua NF-A14x25 G2 PWM Sx2-PP 140mm	
Aux fans	Noctua NF-A4x20 PWM (2-pack)	
Enclosure	Custom 3D-printed acoustic housing & base	

Installed Models

Embedding

Qwen3-Embedding-4B — AWQ INT4

Text embedding and retrieval.

Qwen3.8-27B

27B · Qwen · GGUF Q4_K_M · High-end. The high-end default: strongest general reasoning and agentic tool-use, tuned for everyday workflows.

OLMo 3.1 32B Think

32B · Ai2 · GGUF Q4_K_S · Open data flow. Fully open training data and weights, for teams who want transparent provenance and deep, deliberate local reasoning (Groeneveld et al., 2024).

Both engines run locally and are interchangeable per workspace or per task.

Capability

Runs Spaces in its complete version, with Adaptive mode and full agentic capabilities included: end-to-end local inference and tool use.

Empathy.AI Spaces: Advantages

Empathy.AI Spaces is Empathy.AI's flagship product: a private, self-hosted retrieval-augmented generation (RAG) platform designed to transform how organisations discover, manage and interact with their internal knowledge.

What Spaces Does

Spaces turns fragmented, siloed organisational knowledge, Confluence pages, GitHub repositories, PDFs, DOCX, Markdown files, internal APIs and databases, into a unified conversational AI system. Staff can ask natural-language questions, get accurate, cited answers pulled directly from their own data, and explore insight across documents and workflows without leaving their secure environment.

Q: Which evidence shows competency stagnation in the last term?

A: Three students show a sustained drop in logical-reasoning self-assessments between September and November, coinciding with reduced participation in optional challenges.

 Self_Assessments_Y9.csv  Teacher_Observations_Q1.pdf

Illustrative rendering of Spaces' grounded, cited response pattern, applied to an educational domain.

How It Works: Contextual Retrieval

Spaces uses a technique called **contextual retrieval**, which enriches every document chunk with surrounding context before indexing. This dramatically improves accuracy, reducing retrieval failures by up to **67%** compared to standard chunking methods when combined with reranking (Anthropic, 2024), meaning fewer hallucinations, more relevant answers, and higher trust in AI outputs.

Deployment Options

Option	Description	Best for
On your own infrastructure	Spaces runs on the institution's own server or compute. Empathy.AI provides full turnkey implementation, exactly the scenario described in this blueprint with the Local AI Workstation.	Institutions with compatible compute and a need for full on-premises data sovereignty. Lowest long-term cost.
Leased hardware	High-performance GPU hardware leased from Empathy.AI, hosted in its net-zero facility in Asturias.	Institutions that want full control without owning or managing hardware.
Managed Private Space	Fully managed cloud environment, hosted on Empathy.AI's private, solar-powered, EU-based infrastructure.	Institutions wanting the fastest deployment with zero infrastructure burden.

All options run on open-source LLMs, process data locally, and are aligned with GDPR, the EU AI Act and data-residency law.

Ideal Use Cases

Education · student evidence and minors' data

Legal · confidential client documents

Finance · sensitive transaction data

Healthcare · patient records

Government · strict data-residency rules

Publishing · intellectual property protection

In summary: Empathy.AI Spaces is a private, self-hosted, open-source RAG platform that turns organisational knowledge into conversational AI, never trains models on client data, runs on renewable energy, and is transparent, auditable and built for trust. It is not a black box trained on the internet, but a trusted partner in service of an institution's mission rather than a platform's growth.

Case Study: Putney High School

Running AI on-campus, under school governance. Client: Putney High School (Girls' Day School Trust). Solution: Empathy.AI Spaces on a dedicated Local AI Workstation. Outcome: a privately hosted, sovereign AI environment integrated into the educational experience, with safe AI introduction at Year 9, and student-built, programmatic use at Year 13.

1. At a Glance

Organisation	Putney High School, a selective all-girls school in the GDST family (24 schools, ~100,000 alumnae)
Need	A school-authored way to meet AI head-on in the classroom; computing is timetabled from Year 7, Python from Year 9, and A-Level Computer Science is offered, but “AI appears neither as a standalone offering nor as named policy” (Made in Putney)
Deployment	Empathy.AI Spaces, full version (Adaptive mode plus agentic tool-use), running end-to-end on a Local AI Workstation physically inside the school
Model	Year 9: guided, safe first contact. Year 13: programmatic access; students build new experiences on the same infrastructure
Data posture	Documents, prompts, retrieved evidence, embeddings, model outputs and the reasoning loop remain in the building. Nothing is sent to a public AI API

2. Context: A School Already Built for This Problem

Putney High's stated purpose is educating girls “to be well rounded, well qualified and well prepared,” characterised by “fearless curiosity, intellectual agility, and a joyful ambition for originality and adventure” (Made in Putney). Its technology spine is already drawn: computing from Year 7; Python coding as a third-language option from Year 9 alongside Italian and Spanish; Design Thinking for all Year 9 students; Computer Science to GCSE and A Level. The GDST Girls' Futures Report sharpens the mandate: 54% of surveyed girls follow current affairs via social media, only 28% trust what they see there, and 56% encounter fake news weekly, making judgement, not tools, the priority skill (GDST, *Girls' Futures Report*). The school's own materials record the gap honestly: no specific AI programme yet detailed. The Empathy.AI installation fills exactly that slot: an AI the school governs, rather than a vendor accessed through a corridor.

3. The Solution: An Owned Execution Layer

3.1 Hardware basis

Component	Specification
GPU	NVIDIA RTX PRO 4500 Blackwell (PNY): 32GB VRAM, 896 GB/s memory bandwidth, vertical riser mount
CPU / RAM	AMD Ryzen 7 9700X · Kingston FURY Beast DDR5 32GB (2×16GB)
Motherboard / storage	ASUS ROG STRIX X870-I Gaming WiFi (Mini-ITX) · Kingston NV3 1TB NVMe
Power / cooling	Cooler Master GX II Gold 850W · Noctua low-profile and PWM fan set, custom 3D-printed acoustic housing

Designated capacity: 300W whole-system active draw (~€302–343/month local operating cost at €0.30/kWh); planning bands, not throughput guarantees.