This article has been accepted for publication in IEEE Technology and Society Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MTS.2026.3652471



# Framing the Machine: Teaching Children to Question AI's Voice

**Savvas A. Chatzichristofis**
Neapolis University Pafos, Department of Computer Science, Intelligent Systems Laboratory, Neapolis University Pafos, 8015, Paphos, Cyprus,
Member of the National AI Taskforce of the Republic of Cyprus

***Abstract*—Children are already encountering AI-generated answers in their learning journeys. The critical question is whether these encounters are framed with pedagogical care, ethical oversight, and developmental awareness, or left to the fluency of systems optimized for prediction, not truth. This column argues for structural safeguards in educational AI: tools, interfaces, and delays that prioritize human interpretation over algorithmic authority. At stake is not simply what AI says, but how we choose to hear it, and how we prepare learners to respond. We therefore advocate educator-governed interaction layers that mediate between learners and model outputs by operationalizing delays, human oversight, and auditability; one such reference implementation has been piloted in primary classrooms.**

**When my eight-year-old daughter** asked me who Nikos Kazantzakis was, I didn't recite his biography. I opened ChatGPT, and she watched as I typed. The answer that came back was fluent, polite, and deeply misleading. Kazantzakis, it claimed, was a "20th-century novelist known for his romantic tales and political thrillers," yet the anguish, the spiritual depth, and the existential struggle that shaped his voice were gone. My daughter blinked and said, "That doesn't sound very interesting." This moment did not make me fear artificial intelligence; it reminded me what real education demands: mediation, framing, and human presence. So we rewrote the answer together, keeping the parts she understood and rephrasing the rest; we asked the AI again and compared; we drew his village, listened to Cretan music, and imagined what a conversation with Alexis Zorbas might feel like. Then we wrote our own simplified paragraph. The process took 40 minutes and was the most meaningful lesson of our week.

In what follows, I bring together four complementary perspectives on generative AI in education. I ground the argument in everyday family and classroom experiences that illustrate how children already

*Digital Object Identifier 10.1109/TSM.2025.Doi Number*
*Date of publication DD MM YYYY; date of current version DD MM YYYY*

encounter AI in their learning lives. I also examine technical and architectural choices, such as educator-governed interaction layers, that shape how model outputs are delivered to learners. In addition, I connect these choices to pedagogical frameworks of scaffolding, mediation, and teacher–student co-regulation, and I situate the discussion within emerging regulatory and policy developments, including the EU AI Act and international guidance from UNESCO and the OECD. Together, these perspectives help clarify not only what AI says, but how we choose to frame what children hear.

For the classroom, I've seen the same vision. Generative AI is not an oracle or a tutor; it is a mirror and a provocateur. In a recent project [1], we explored how large language models could be used to simplify canonical literary texts for younger readers, such as passages from The Tempest and Macbeth. The system generated age-aligned versions, adapting syntax, vocabulary, and emotional tone while preserving semantic depth. These outputs were not final answers but starting points for reflection. Students and educators examined parallel versions of the texts, questioned lexical choices, and reflected on tone, affect, and audience. The goal was not to let AI decide, but to support interpretation through guided comparison and discussion. With appropriate scaffolding, learners engaged critically, as editors, interpreters, and co-constructors of meaning. Our emphasis is human-centered: teacher judgment, dialogue, and community norms frame the work; technology is a resource shaped by those relationships, not an end in itself.

In addition to principle-level guidance, UNESCO has also published AI competency frameworks for students and teachers [3], [4], which we include here to connect public governance directly to classroom-facing competencies. We treat these frameworks as competency targets; the concrete interface and governance mechanisms described below are our proposed operationalization to support those targets in classroom settings.

Generative tools can support rich learning when carefully scaffolded; without such framing, the picture changes. A recent meta-analysis by Wang and Fan [7] reports gains in learning, perception, and higher-order thinking, but only under appropriate scaffolds. Without scaffolding, children are exposed to systemic risks: hallucinated facts, incoherent reasoning, and the seductive confidence of statistically plausible language. AI can dazzle without informing, affirm without explaining, and simplify without understanding. The problem is not that AI makes mistakes; the problem is that its voice sounds right, especially to a child. When children receive unframed responses, they may mistake fluency for accuracy, or correlation for meaning. The AI's voice becomes dominant, yet although some systems now attach citations or generate step-by-step explanations, recent evidence indicates these self-explanations are often not faithful to the model's actual decision process and may mislead non-experts [8], so it cannot reliably explain its reasoning or justify its claims without educator mediation.

This is why no child should ever face generative AI alone: the solution is not prohibition, but thoughtful design. We need transparent intermediary layers, software governed by educators, ministries, and communities, that filter and contextualize AI responses for children. This aligns with growing calls for public oversight of AI tools in education, such as those embedded in the EU AI Act's classification of educational systems as high-risk. These interfaces should align outputs with ethical guidelines (e.g., UNESCO's 2021 Recommendation on the Ethics of Artificial Intelligence [2], UNESCO's AI competency frameworks for students and teachers [3], [4], the OECD AI Principles [5], and the OECD and European Commission's AI Literacy Framework [6]), support curricular goals, and scaffold interpretation. This emphasis on pedagogical framing is consistent with our related argument that AI's educational value depends on ethical framing and educator-led embedding, rather than on technical affordances alone [9].

In Table 1, the mappings indicate how an educator-governed interaction layer can help operationalize these public frameworks in practice, rather than implying that the frameworks prescribe specific technical features. Their purpose is not to suppress AI tools, but to reframe them in ways that are developmentally, pedagogically, and culturally coherent. This is a call for public institutions to establish design principles before private tools dominate the educational space by default. To make this explicit, we locate key decisions in public institutions: curriculum authorities, school boards, and parent councils set defaults, approve prompt libraries, review audit logs, and publish clear opt-out routes, while professional associations and civil society monitor impacts on equity, privacy, and labor. More broadly, UNESCO's 2025 capacity-building work on AI supervisory authorities underscores that effective governance depends on practical

This article has been accepted for publication in IEEE Technology and Society Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MTS.2026.3652471

**TABLE 1. Public frameworks mapped to educator-governed interaction layers in schools**

| Document | Issuer | Year / Status | Scope | Relevance to an educator-governed interaction layer and AI literacy coverage |
|---|---|---|---|---|
| UNESCO *Recommendation on the Ethics of AI* [2] | UNESCO | 2021 (adopted) | Ethical principles and safeguards | Human rights, transparency, accountability, protection of children, human oversight. In line with UNESCO's 2023 guidance on generative AI in education, countries are recommended to set a minimum age for classroom use (often around the age of digital consent, e.g., 13) with strong mediation; an educator-governed layer enables mediated or indirect exposure with teacher pre-review and logging. |
| UNESCO *AI competency framework for students* [3] | UNESCO | 2024 (published) | K–12 student competencies | Classroom-facing competencies for students in the education context; an educator-governed interaction layer *can support* these through institution-defined scaffolds (e.g., staged disclosure, reflective prompts, and logged inquiry cycles) under teacher oversight. |
| UNESCO *AI competency framework for teachers* [4] | UNESCO | 2024 (published) | Teacher competencies and professional learning | Teacher-focused competencies that emphasize human-centred, ethical, and pedagogical AI use; an educator-governed interaction layer *can support* teacher oversight and decision-making through institution-defined controls (e.g., review gates, role-based permissions, and audit logs) for teacher-led, documented classroom use. |
| OECD *AI Principles* (Council Recommendation on AI) [5] | OECD Council | 2019 (OECD/LEGAL/0449) | Values and policy guidance | The layer *supports the operationalization* of core principles (transparency, accountability, robustness, safety) via deployer instructions, auditable logs, provenance, staged human oversight and risk-oriented governance—consistent with the Recommendation's implementation guidance. |
| *AI Literacy Framework for Primary and Secondary Education* [6] | OECD and European Commission | 2025 (review draft) | K–12 AI literacy domains | *Organized into four domains*—**Engaging with AI**, **Creating with AI**, **Managing AI**, **Designing AI**. The layer aligns as follows: *Engaging with AI*: deliberate disclosure delay, teacher pre-review, reflective prompts, age-tiered defaults; *Creating with AI*: staged disclosure (cues → exemplars → full answers), provenance/versioning, tri-modal outputs (student text / teacher notes / reflections); *Managing AI*: school-level guardrails, data minimization, dashboards, role-based permissions, configurable time-to-disclosure; *Designing AI*: sandboxed explorations, scenario-based investigations, model comparison, logged inquiry cycles with responsible defaults. |

institutional routines and supervisory capacity that translate high-level norms into operational practice [10]. Concretely, under Regulation (EU) 2024/1689, AI used to determine access or admission, evaluate learning outcomes, assign individuals to levels, or monitor exams is classified as high-risk (Annex III, 3).

A promising example solution to these issues is an educator-governed validation layer (e.g., Ethical–Pedagogical Validation Layer (EPVL) [1]). This middleware sits between the model and the learner and introduces a structured space for ethical and pedagogical judgment. Rather than offering raw outputs, the layer produces three views: a simplified student-facing text, pedagogical commentary for educators, and reflective prompts to stimulate critical engagement. As illustrated in Figure 1, student queries pass through a school-governed interaction layer that applies public guidelines and teacher review before any model output reaches the classroom, while all interactions are logged for accountability.

From a regulatory and policy perspective, this interaction-layer pattern supports alignment with key EU AI Act requirements by embedding human oversight before disclosure (Article 14), providing institution-facing instructions and disclosures (Article 13), enabling lifecycle logging for auditability (Article 12), and scaffolding risk management and data governance processes (Articles 9–10) [11], [12].

From a technical and architectural perspective, the layer acts as adaptable middleware between the model and the learner. It should be deployed as institution-

**TABLE 2. EU AI Act requirements mapped to an educator-governed interaction layer**

| AI Act reference | What it requires | How the layer addresses it |
|---|---|---|
| Article 9: Risk management system | Risk identification, analysis, mitigation across lifecycle | Policy-as-code guardrails, configurable thresholds, and governance workflows that document risks and mitigations; linkage of prompts/outputs to risk registers and approval trails. |
| Article 10: Data governance | Data quality, relevance, representativeness; data governance measures | Data minimization and residency controls; role-based access; retention schedules; provenance metadata captured in immutable logs to support audits and reviews. |
| Article 12: Record-keeping and logging | Comprehensive technical logs for traceability and audits | Versioned, immutable audit logs (prompts, model identifiers, outputs, educator actions/approvals, timestamps, policy decisions), with retention and access controls set by institutional policy. |
| Article 13: Transparency & information to users | Providers must supply instructions for use, capabilities/limits, intended purpose | The layer surfaces provider documentation to institutional administrators and binds intended purpose and capability limits to configuration defaults and on-screen admin guidance. |
| Article 14: Human oversight | Effective oversight, ability to intervene/override/suspend | Pre-disclosure hold, teacher approval gates, pause/override controls, escalation routes; every intervention is logged and reviewable. |
| Article 50: Transparency duties for certain AI systems | Users must be informed when interacting with AI (where applicable) | User-facing notices and indicators when model-generated content is shown; contextual disclosures aligned with Article 50 obligations. |
| Annex III.3: High-risk uses in education | Admission, grading/assessment, placement, or monitoring/evaluating examinations are high-risk | Administrative defaults: role-based access, curriculum tagging, age-tiered settings, logging on by default, auditable workflows, and clear routes for concerns and redress under public-authority and school policies. |
| Article 27: Fundamental rights impact assessment (FRIA) | FRIA for public authorities or providers of public services using high-risk AI | The layer produces evidence packs (usage analytics, logs, policy configs, DPIA links) to support FRIA preparation and periodic reviews by institutions and public authorities. |

controlled middleware with on-premises or regionally hosted inference endpoints behind an auditable API gateway that enforces pre-disclosure validation, policy-as-code guardrails, and data residency constraints. Core functions should run as containerized microservices with role-based access control, versioned artifact storage, immutable and queryable logs, and encryption in transit and at rest to ensure provenance and full classroom oversight. Integration should follow model-agnostic APIs so schools can switch between commercial services and locally hosted open models without modifying the pedagogical layer, while defaulting to strict data minimization, zero data retention unless opted in, and end-to-end logging.

From a pedagogical and learning sciences perspective, its true value is that it delays the presentation of AI-generated responses until a human educator can first review, contextualize, or reshape them. This intentional pause enables teachers to scaffold student reasoning before the model's answer appears, drawing on principles from Vygotsky's Zone of Proximal Development [13] and Schön's theory of reflective practice [14]. Accordingly, educator agency and classroom context determine when and how the system is used; the layer simply structures the interaction so that mediation remains in human hands. In this article, we use *scaffolding* to mean contingent, adaptive support that fades as learner competence grows (aligned with the ZPD), *mediation* to mean the cultural and semiotic tools through which meaning is constructed (e.g., teacher prompts, rubrics, AI-generated drafts), and *additional layer* to denote the infrastructural interaction layer (this educator-governed layer, e.g., EPVL) that operationalizes mediation and enables teacher-led scaffolding at the point of use [13], [15], [16], [17]. We make the target of scaffolding explicit:

This article has been accepted for publication in IEEE Technology and Society Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MTS.2026.3652471

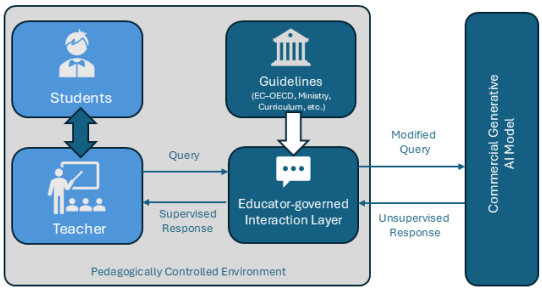


**FIGURE 1.** Educator-governed interaction layer in a pedagogically controlled environment. Student queries are mediated by teacher review and policy guardrails that implement public guidelines (UNESCO, OECD, Ministry, etc.), with provenance and logging, before any response from the commercial model is shown.

(a) teacher-facing scaffolding comprises pedagogical commentary, risk flags, and suggested prompts that support planning and in-situ mediation; (b) student-facing scaffolding comprises reflective questions, partial exemplars, and age-tiered explanations designed to fade as competence increases; and (c) joint scaffolding combines both through co-regulation cycles in which teachers set constraints and timing while the system surfaces cues and students act on them [17], [18]. We distinguish scaffolding from tool-centric assistance that accelerates production without fading support (e.g., coding copilots that tend to supply solutions by default); in our approach, support aims at independence rather than offloading. This design aligns with the augmentation perspective and with hybrid human–AI learning technologies that articulate teacher–AI co-regulation, detect–diagnose–act cycles, and levels of automation [19], [18]. To address specific weak points in child–AI interactions, the interaction layer implements or supports pre-disclosure human review to prevent premature anchoring, curriculum tagging and age-tiered simplification to align outputs with learning goals, source disclosure and versioned logging to counter hallucinations, reflective prompts to sustain metacognitive questioning, and accessibility options for multilingual and diverse learners.

From an everyday classroom perspective, in pilots with learners aged 10 to 12 [1], students began to treat AI as a draft generator rather than an authoritative source, asking, "Why did it say this?" instead of simply copying the response. This shift from passive use to dialogical engagement reflects the power of interpretive mediation, and the architecture of such a layer enables real-time educator feedback that supports continuous system refinement and alignment with national standards.

This layer is not a product; it is a governance pattern in technical form, a replicable blueprint for jurisdictions seeking to retain pedagogical sovereignty in ecosystems dominated by commercial AI interfaces. It empowers public institutions to define what "responsible AI in education" looks like in practice, by embedding ethical, cultural, and developmental priorities directly into the interaction layer.

We use EPVL as a reference pattern of an educator-governed interaction layer. See Table 1 for how public frameworks map to requirements that such a layer can operationalize. Concretely, the educator-governed interaction layer turns public guidance into operations: pre-disclosure teacher review, role-based permissions, and immutable audit logs make transparency, accountability, and human oversight routine classroom practices, as required by UNESCO's *Recommendation on the Ethics of AI* and consistent with the OECD *AI Principles*. Provenance indicators, versioning, age-tiered defaults and refusal routes uphold child protection and responsible disclosure emphasized by UNESCO. For primary and secondary AI literacy, the layer supports *Engage* and *Manage* through configurable gating and logging; it enables *Create* and *Design* through staged disclosure (prompts and cues first, exemplars next, full solutions only when pedagogically justified) and sandboxed exploration, meaning a school-controlled safe environment where students can experiment with prompts and compare outputs under teacher oversight and logging, that keeps authority with teachers, aligning with the OECD and European Commission *AI Literacy Framework*.

Table 2 summarizes how the educator-governed interaction layer operationalizes key EU AI Act provisions. Concretely, it implements versioned, immutable logs that capture prompts, model identifiers, outputs, educator actions, timestamps, and policy decisions to satisfy record-keeping under Article 12; provides inline disclosures of intended purpose, capability limits, provenance, and age-tiered defaults to meet transparency under Article 13, with user-facing indicators when interacting with AI where applicable (Article 50); and enforces pre-disclosure holds, teacher-approval gates, pause/override controls, and escalation routes to deliver human oversight under Article 14. For Annex III.3 high-risk educational tasks, role-

This article has been accepted for publication in IEEE Technology and Society Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MTS.2026.3652471


based access, curriculum tagging, default-on logging, auditable workflows, and clear routes for concerns and redress keep schools and public authorities in effective control. The layer also supports Articles 9–10 through policy-as-code guardrails, data minimization, retention schedules, and provenance metadata, and assembles evidence packs (usage analytics, logs, policy configurations, DPIA links) to facilitate the fundamental-rights impact assessment under Article 27. In short, it lets schools apply the Act as routine classroom practice rather than ad hoc compliance.

In Estonia, where AI tutors are integrated into national systems, such scaffolding is already visible, while in Cyprus and other Mediterranean countries local initiatives are beginning to emerge. But these efforts are scattered. The private platforms move faster and reach our homes before our ministries do. So the question is no longer whether children will hear from AI, because they already do; the real question is whether what they hear will be framed with care, context, and human intention, or left to the indifferent fluency of an algorithm optimized for prediction, not truth. Kazantzakis once wrote, "You have your brush, you have your colors, you paint the paradise, then in you go." Let us ensure that the brushes we give our children are governed by pedagogical values, framed by reflective practice, and informed by human understanding. Let us not offer them just fluency, but wisdom. Not just power, but trustworthiness. Not just mirrors, but companions for thought.

This requires more than ethical aspirations; it demands concrete, structural safeguards. We need certified educational interfaces that foreground pedagogical purpose over performance, and we need public prompt libraries that are transparently designed, aligned with curricular goals, and accessible to all educators. We also need deliberate response delays that open space for teacher-led scaffolding before any AI answer appears, as well as reflective AI kits that let students explore variations, revise suggestions, and interrogate alternatives. Even where vendors advertise "educational" or "tutor" modes that support iterative, critical questioning, these remain optional interface features rather than enforceable, educator-governed controls; our proposal specifies institutionally set delays, age-tiered defaults, and auditable mediation at the interaction layer. Operationally, staged disclosure is used, prompts and cues first, exemplars next, and full solutions only when pedagogically justified, with control shifting to learners over time and dashboards supporting staged transfer of control; the process is logged so that teachers can withdraw supports as proficiency grows [18].

In parallel, educators need protected time and funded professional development to design prompts, curate examples, and conduct classroom-level reviews, together with school policies that recognize refusal rights and protect teacher autonomy.

In short, specifically designed technology is necessary but not sufficient: educators lead classroom use, school leaders and parent councils provide oversight, public authorities set defaults and standards, and vendors adapt to these constraints.

Finally, we need audit trails for educational AI outputs, including versioning, source attribution, and disclosure of model provenance. These logs directly address the Act's record-keeping obligation for high-risk systems and facilitate the institutions' and public authorities' fundamental-rights impact assessment (Articles 12 and 27).

Technically, these safeguards can be modular and scalable, whereas pedagogically they are indispensable. Systems that adopt an educator-governed interaction layer, with delayed disclosure and auditable mediation, offer a promising blueprint: less a final answer than a direction that keeps interpretive sovereignty with learners and teachers.

## REFERENCES

1. S. A. Chatzichristofis, A. Tsopozidis, A. Kyriakidou-Zacharoudiou, S. Evripidou, and A. Amanatiadis, "Designing an ai-supported framework for literary text adaptation in primary classrooms," *AI*, vol. 6, no. 7, p. 150, 2025.
2. UNESCO, "Recommendation on the ethics of artificial intelligence," 2021, adopted by UNESCO Member States, 23 Nov 2021. Official text available via UN Digital Library. [Online]. Available: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence
3. F. Miao, K. Shiohira, and N. Lao, *AI competency framework for students*. Paris, France: UNESCO, 2024. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000391105
4. F. Miao and M. Cukurova, *AI competency framework for teachers*. Paris, France: UNESCO, 2024. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000391104
5. OECD, "Recommendation of the council on artificial

intelligence," OECD Legal Instrument OECD/LEGAL/0449, 2019. [Online]. Available: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

6. European Commission and OECD, "Ai literacy framework for primary and secondary education," Review Draft, 2025, joint EC–OECD initiative; public consultation draft. [Online]. Available: https://ailiteracyframework.org/wp-content/uploads/2025/05/AILitFramework_ReviewDraft.pdf
7. J. Wang and W. Fan, "The effect of chatgpt on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis," *Humanities and Social Sciences Communications*, vol. 12, no. 1, pp. 1–21, 2025.
8. A. Madsen, S. Chandar, and S. Reddy, "Are self-explanations from large language models faithful?" in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. [Online]. Available: https://aclanthology.org/2024.findings-acl.19/
9. S. A. Chatzichristofis, "Reframing ai in education: A pedagogical opportunity, not a technocratic threat," *Journal of Research in Science Teaching*, vol. 63, pp. 97–103, 2026. [Online]. Available: https://doi.org/10.1002/tea.70024
10. UNESCO, "Pathways on capacity building for ai supervisory authorities: Insights and recommendations from the 1st unesco expert roundtable on ai supervision," Paris, 2025.
11. "Regulation (eu) 2024/1689 — artificial intelligence act," https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng, 2024, official Journal of the European Union.
12. "Eu ai act explorer — section 2: Requirements for high-risk ai systems," https://artificialintelligenceact.eu/section/3-2/, 2024.
13. L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*, M. Cole, V. John-Steiner, S. Scribner, and E. Souberman, Eds. Cambridge, MA: Harvard University Press, 1978.
14. D. A. Schön, *The Reflective Practitioner: How Professionals Think in Action*, 1st ed. London: Routledge, 1992, original work published 1983. [Online]. Available: https://doi.org/10.4324/9781315237473
15. D. Wood, J. S. Bruner, and G. Ross, "The role of tutoring in problem solving," *Journal of Child Psychology and Psychiatry*, vol. 17, no. 2, pp. 89–100, 1976.
16. S. Puntambekar and R. Hubscher, "Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed?" *Educational Psychologist*, vol. 40, no. 1, pp. 1–12, 2005.
17. J. J. G. van Merriënboer and P. A. Kirschner, *Ten Steps to Complex Learning: A Systematic Approach to Four-Component Instructional Design*, 3rd ed. New York, NY: Routledge, 2018.
18. I. Molenaar, "The concept of hybrid human-ai regulation: Exemplifying how to support young learners' self-regulated learning," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100070, 2022.
19. ——, "Towards hybrid human-ai learning technologies," *European Journal of Education*, vol. 57, no. 4, pp. 632–645, 2022.