

Teaching With, Not Around

A Developmental Framework for Generative AI Integration in Primary, Secondary, and Tertiary Education (2023–2026)

N.J. Berky

NJB Berky Advisory · Prague / Zürich

ORCID: 0009-0001-2227-215X · njb@njb-berky.com · njb-berky.com

Working paper · Version 2.0 · June 2026

ABSTRACT

Generative artificial intelligence (GenAI) is the fastest-diffusing general-purpose technology in modern educational history. Between November 2022 and early 2026, frontier large language models moved from research curiosity to ubiquitous student tool in three academic cycles. Education systems did not. This paper argues that the dominant institutional postures, prohibition, unregulated adoption, and surveillance-based compliance, are all developmentally illiterate, and proposes a four-stage cognitive framework (Delegation, Description, Discernment, Diligence; hereafter 4D) calibrated against the neurodevelopmental realities of primary, secondary, and tertiary learners. Drawing on a synthesis of empirical work from 2023 to 2026 on cognitive offloading (Kosmyrna et al., 2025; Gerlich, 2025; Fan et al., 2025; Gilbert, 2024), the developmental neuroscience of prefrontal maturation under digital exposure (Hutton et al., 2022; Hinze et al., 2025; Lewin et al., 2025; Gascón-Cánovas et al., 2025), the EU and OECD AI literacy frameworks (European Commission & OECD, 2025; UNESCO, 2024), and the forthcoming PISA 2029 Media and AI Literacy assessment (OECD, 2026), the paper makes four claims. First, GenAI exposure should be structurally asymmetric across developmental tiers, not uniformly permitted or uniformly forbidden. Second, model selection in regulated educational environments is a non-trivial pedagogical choice; methodology-transparent systems such as those trained under Constitutional AI protocols (Bai et al., 2022) present a more defensible foundation for institutional deployment than systems where alignment methodology is undisclosed. Third, the educator's role is shifting from content-deliverer to cognitive coach, and educator training must shift correspondingly. Fourth, the international policy environment is converging rapidly around four-domain literacy models (engage, create, manage, design) which create both an obligation and an opportunity for institutional reform. The paper concludes with operational guidance for institutions across the 2026–2030 horizon, a limitations section, and an agenda for future empirical research.

Keywords: generative AI · AI literacy · cognitive offloading · developmental neuroscience · educational technology · 4D framework · Constitutional AI · AI ethics in education · adolescent executive function · OECD AILit framework · PISA 2029 · protection-by-design pedagogy

1. Introduction · The Asymmetry Problem

In the eighteen months between OpenAI's public release of ChatGPT (November 2022) and the broad institutional adoption of frontier large language models (LLMs) by professional services, software development, and research workflows (2024), generative AI completed a market-penetration arc that earlier general-purpose technologies (electricity, the internal combustion engine, the personal computer, the internet) required decades to achieve. By early 2026, frontier models such as Claude (Anthropic), GPT-4 and GPT-5 (OpenAI), and Gemini (Google DeepMind) operate as default reasoning and writing partners for tens of millions of students worldwide, including across primary education systems whose curricula were not designed to accommodate them. A recent meta-analysis of 35 controlled studies involving 4,193 participants reports a moderate positive effect of

ChatGPT on student learning outcomes (Hedges' $g = 0.670$) across primary, secondary, and tertiary contexts (Deng et al., 2026), confirming that the technology is not pedagogically inert. The question is no longer whether students use these systems. They do. The question is whether they use them in ways that build cognitive capacity or substitute for it.

This is not a tooling problem. It is a developmental one. A nine-year-old, a fifteen-year-old, and a twenty-two-year-old engaging with the same large language model are not engaging with the same artefact; their prefrontal cortices, working-memory capacities, metacognitive reflexes, and epistemic defaults are radically different. The prefrontal cortex, the seat of executive function, judgement, and inhibitory control, does not reach mature myelination until the mid-twenties (Diamond, 2013). The cognitive routines that allow an adult researcher to use a frontier model constructively, source verification, premise interrogation, output disagreement, are precisely the routines a primary-school child is in the active process of forming. Treating GenAI as a uniform classroom technology, to be permitted, restricted, or audited at a single institutional threshold, is a category error. The brain that uses the tool determines what the tool does to the learner.

Educational systems globally have responded with three dominant postures, each, in my view, inadequate. The first is prohibition: outright bans or restrictions on student GenAI use, currently in force across roughly forty percent of national education systems in some form (UNESCO, 2024). Prohibition is enforcement theatre. It cannot survive contact with a generation that carries frontier models in its pockets, and it produces a perverse outcome: students learn to use AI without supervision, without instruction, and without scaffolding precisely because institutional silence pushes their learning into the shadow curriculum (Kharbach, 2026). The second is unregulated adoption: institutional silence permitting students to integrate GenAI into every cognitive task without scaffolding, supervision, or pedagogical framing. This is the modal posture of most tertiary institutions today, and it produces what Fan et al. (2025) describe as metacognitive laziness, the systematic erosion of the self-regulatory routines that distinguish learning from completion. The third is surveillance-based compliance: deploying AI-detection tools (which do not reliably work) and conducting disciplinary processes against students for AI use that institutions have failed to teach them how to navigate. This posture is corrosive to institutional legitimacy and produces neither learning nor compliance.

All three postures share a single failure: they treat GenAI as an external object to be controlled, rather than as a cognitive partner whose appropriate use is a developmentally-staged skill. This paper argues the opposite. AI fluency, like literacy, numeracy, or critical reasoning, is a competence that must be deliberately taught, sequentially, calibrated to the cognitive capacities of the learner at each developmental tier. The argument is not that GenAI is inherently good or inherently bad for cognition. It is that the question is malformed. The relevant questions are: for which cognitive tasks, at which developmental stage, under what supervision, with what protection, for what pedagogical purpose. These are questions of pedagogy, not policy alone.

I propose a four-stage cognitive framework (the 4D framework: Delegation, Description, Discernment, Diligence) and demonstrate how each stage should be introduced, sequenced, and reinforced across primary (ages 5–11), secondary (12–17), and tertiary (18+) education. I draw on empirical work from 2023–2026 on cognitive offloading and metacognitive laziness, developmental neuroscience of executive function under digital exposure, and the convergent AI literacy frameworks now emerging from UNESCO (2024), the European Commission and OECD (2025), and national education authorities. I

argue, finally, that model selection itself is a pedagogical choice: institutions deploying GenAI to minors carry a fiduciary obligation to select systems whose alignment methodology is transparent and auditable, and I examine the Constitutional AI methodology developed by Anthropic (Bai et al., 2022) as one defensible foundation for such selection.

This paper is offered as a working framework, not a finished one. Empirical evidence on the cognitive consequences of sustained GenAI exposure is, as of mid-2026, accumulating rapidly but is far from complete. The recommendations here represent my best synthesis of available evidence combined with operational experience teaching applied AI at the tertiary level across IFSA Prague, CEA CAPA, and Anglo-American University Prague, mentoring early-stage ventures through public-private innovation hubs in Central Europe, and architecting deployed multi-paradigm AI systems for regulated industry contexts. Section 12 lays out the limitations of the present analysis and an agenda for the empirical work the framework requires.

2. The Acceleration Problem · AI Velocity and Educational Response

Between GPT-3.5's public release in November 2022 and the deployment of frontier reasoning models in early 2026, the capability frontier of publicly accessible LLMs advanced through at least four major generational shifts. Standardised benchmarks demonstrate compound improvement: models that produced incoherent paragraphs in 2022 routinely produce graduate-level analytical writing in 2026; models that could not perform multi-step mathematical reasoning in 2023 now solve competition-level problems; models that hallucinated basic facts in 2024 are increasingly capable of citing real, verifiable sources. The trajectory is not linear. It is, by any reasonable measure, exponential within the 2022-2026 window. The doubling time for capabilities on key reasoning benchmarks has remained shorter than two years across the period, and inference costs have fallen by approximately an order of magnitude per year, making frontier capability accessible to ever-younger and ever-poorer learners.

Institutional educational systems are not exponential. They are, by design, conservative. Slow to update curricula, slow to retrain teachers, slow to revise assessment frameworks, slow to renegotiate the implicit contract between learner, institution, and credential. This is not a flaw; stability and continuity are core institutional functions of education. The credential system depends on it. So does the trust contract between teacher, student, and the body of knowledge being transmitted. The slowness is the feature. It is, however, a structural mismatch in the current moment. The capability gap between what students can do with frontier AI and what their institutions have taught them to do with frontier AI widens with every model release.

Three concrete consequences follow. First, the shadow curriculum: students are now learning to use GenAI primarily through commercial platforms, social media tutorials, peer transmission, and unsupervised trial-and-error (Kharbach, 2026; UNESCO, 2024). Whatever institutions formally teach about AI is consistently a generation behind what students informally encounter. The shadow curriculum is unregulated, uneven by socioeconomic status, and pedagogically variable. It is also, increasingly, the primary mechanism by which children form their conception of what AI is and how it should be used. Second, the assessment-validity collapse: traditional homework, take-home essay, and many forms of in-class written assessment no longer reliably measure individual cognitive performance, because the boundary between student work and AI-assisted work

is no longer detectable with confidence. Detection tools have proven unreliable, particularly for non-native English speakers, on whom they produce systematically higher false-positive rates (Liang et al., 2023, as reviewed in Kharbach, 2026). Third, the credential-trust problem: employers, regulators, and downstream institutions increasingly cannot infer from a credential what a learner can actually do unassisted. This is a slow-moving but potentially severe institutional risk for the credential system itself.

Two policy developments in 2024–2025 are worth foregrounding. The European Union's AI Act entered into force on 1 August 2024, with Article 4 imposing a binding AI literacy obligation on providers and deployers of AI systems with phased implementation across 2025–2027 (European Union, 2024). UNESCO released two AI competency frameworks (one for students and one for teachers) at Digital Learning Week 2024, defining 15 teacher competencies across five dimensions and a parallel framework for students (UNESCO, 2024; Miao & Cukurova, 2024). The European Commission and OECD jointly released the AI Literacy Framework for Primary and Secondary Education (AILit) in May 2025, organising AI literacy around four practical domains (engaging with AI, creating with AI, managing AI, designing AI) supported by 22 competences (European Commission & OECD, 2025). The OECD published the first draft of the PISA 2029 Media and Artificial Intelligence Literacy (MAIL) assessment framework in February 2026 (OECD, 2026), establishing that AI literacy will be a measured, internationally-comparable competence by the end of the decade. Global readiness against these frameworks has been estimated at under 4% (Welker, 2025), implying roughly 96% of national educational systems are operating without the competency infrastructure their populations now require.

The framework offered in this paper is one possible response to that gap. It is not the only possible response. It is calibrated specifically to the cognitive-developmental realities of K–12 and tertiary learners and to the safety properties of currently available frontier models. Other responses, whether purely structural, purely policy-driven, or purely tool-driven, are possible and worth pursuing in parallel. The 4D framework presented here is intended to be complementary to the AILit four domains, the UNESCO frameworks, and emerging national frameworks; it operates at a different level of abstraction, focusing on the cognitive routines a learner must develop to use AI productively at each developmental stage, rather than on the competency outcomes those routines should produce.

3. Empirical Foundations · A Synthesis of 2023–2026 Evidence

Before proposing a framework, it is necessary to clarify what the empirical literature does and does not establish. The evidence base on GenAI in education is now substantial but uneven; some findings are robust and converging, others are preliminary and contested. This section summarises the state of the evidence as of mid-2026 across four clusters: cognitive offloading and metacognitive effects, learning-outcome studies, neurodevelopmental and behavioural correlates of intensive digital exposure, and the international policy and competency landscape. The framework proposed in subsequent sections is built explicitly on this evidence base, with the strength of each evidentiary claim noted where it bears on a recommendation.

3.1 Cognitive offloading and metacognitive effects

Cognitive offloading refers to the use of external resources to reduce internal cognitive demand (Gilbert, 2024). It is, in general, a normal and adaptive feature of human cognition; we offload to notebooks, calendars, search engines, and other people routinely. The empirical question GenAI raises is whether the specific properties of large language

models, fluent natural-language output, low-friction access, immediate validating response, produce patterns of offloading that systematically erode the cognitive routines learners need to develop.

Three studies from 2025 are central to this question. Kosmyna et al. (2025), in a MIT Media Lab technical report, used neurophysiological measurement (EEG-derived cognitive load metrics) to examine essay-writing performance across conditions in which adult participants used an AI assistant, used standard internet search, or used neither. The study reports accumulation of what the authors term cognitive debt, a measurable reduction in engaged cognitive effort during the AI-assisted condition that persisted across sessions and was associated with lower self-reported task engagement. Gerlich (2025), in a published study in *Societies*, reports a parallel finding in a larger but less methodologically intensive sample: higher self-reported AI tool reliance correlated with lower critical-thinking performance on standard instruments, with cognitive offloading mediating the relationship. Fan et al. (2025), in the *British Journal of Educational Technology*, designed an experimental study using essay-writing tasks to investigate what they term metacognitive laziness, the systematic reduction of self-regulatory cognitive activity under AI-assisted conditions, and report consistent effects on learning motivation, processes, and performance.

These studies should be read with appropriate caution. None is a longitudinal study of children. Kosmyna et al. (2025) studied adults at essay-writing tasks; the cognitive routines being measured are mature, not developing. Gerlich (2025) relies on self-report for both AI use and critical-thinking outcomes, with the methodological limitations such designs imply. Fan et al. (2025) measure short-term effects, not developmental consequences. Nonetheless, the convergence across three independent research groups, using different methodologies, on a common finding (that low-friction AI assistance reduces engaged cognitive activity in ways that are detectable both behaviourally and physiologically) is meaningful. It is not yet sufficient to establish that childhood AI exposure causes developmental cognitive harm. It is sufficient to take seriously the hypothesis that pedagogical design which fails to deliberately introduce friction will produce learners who have not built the cognitive routines AI use requires.

Two further studies extend the picture. Ergene and Ergene (2025), in *Education and Information Technologies*, examine cognitive offloading and over-dependence on GenAI specifically in higher education contexts, reporting both productive and dependence-pattern outcomes depending on instructional framing. Lee and Low (2024), in *Computers & Education: Artificial Intelligence*, examine trade-offs between efficiency and originality in academic writing under GenAI assistance, reporting that students producing AI-assisted text complete tasks faster but produce work with measurably lower linguistic and conceptual diversity. The implication is not that AI assistance is categorically harmful for academic writing; it is that the routine that produces originality (sustained personal engagement with a problem) is one of the routines AI assistance most easily substitutes for. Pedagogical design must protect that routine deliberately.

3.2 Learning-outcome studies

A countervailing body of evidence reports positive learning effects from GenAI integration. Deng, Wang, and Zhang (2026), in a meta-analysis of 35 experimental studies covering 4,193 participants published in 2022–2024 (Humanities and Social Sciences Communications), report a moderate positive effect of ChatGPT on student learning outcomes (Hedges' $g = 0.670$), with significant moderating effects of subject area, experimental duration, and instructional mode but not of educational level. Cunningham

et al. (2025), in a year-long study at the University of Tennessee at Chattanooga (Frontiers in Education), report transformative impacts of ChatGPT integration on graduate-student motivation and skill acquisition, with the largest effects in structured-task contexts. Raptopoulou (2025), in a Swedish university qualitative study (European Journal of Education), reports that ChatGPT supports academic literacy development when integrated with clear pedagogical goals but not when deployed without scaffolding.

The apparent tension between the offloading literature (Section 3.1) and the outcomes literature (here) dissolves when the dependent variables are examined carefully. The offloading studies measure cognitive engagement and routine formation, the outcomes studies measure task completion and grade-level performance. It is entirely consistent for AI-assisted students to complete tasks better and learn less, if learning is understood as the formation of cognitive routines rather than the production of correct outputs. This distinction, between task-completion outcomes and routine-formation outcomes, is the central pedagogical distinction the present framework operationalises. Section 4 develops it in detail.

3.3 Neurodevelopmental and behavioural correlates

Direct neurodevelopmental research on GenAI exposure does not yet exist in any methodologically robust form; the technology is too new and the longitudinal timescales required (years to decades) preclude it. The relevant proxy literature concerns digital exposure more generally, particularly the developmental neuroscience of intensive screen and smartphone use in children and adolescents. Hutton et al. (2022), in JAMA Pediatrics, report associations between screen-based media use and reduced white matter integrity in preschool-aged children, specifically in tracts supporting language and emergent literacy. León-Domínguez (2024), in Neuropsychology, reviews potential cognitive risks of generative transformer-based AI chatbots on higher-order executive functions. Lewin et al. (2025), in the International Journal of Clinical and Health Psychology, report longitudinal effects of screen media activities on brain function, architecture, and mental health in early adolescence. Gascón-Cánovas et al. (2025) report differential executive function performance in adolescents with problematic smartphone use compared to matched controls (110 students, ages 13–18), with effects detectable on parent-report measures even where controlled lab tasks underestimate them. Hinze et al. (2025), in a longitudinal study of 8,072 UK adolescents from the MYRIAD trial, demonstrate stable bidirectional associations between adolescent executive function and mental health outcomes including well-being, depressive symptoms, and suicidality.

A scoping review by Reggente et al. (2026), in Pediatric Reports, analyses 104 articles published 2022–2025 on smartphone use and adolescent brain function, identifying five recurring themes: psychological disturbances ($n = 52$), sleep ($n = 43$), socioemotional function ($n = 23$), executive function ($n = 14$), and sensory processing ($n = 1$). The review documents reduced functional connectivity in cortico-subcortical circuits, reduced grey matter volume in superior and inferior frontal gyrus, and aberrant cortical activation in frontoparietal regions during cognitive conflict tasks among adolescents with high smartphone use. The findings are correlational, not causal, and the specific extension to GenAI-mediated digital exposure remains to be empirically established. They are, however, the best available proxy evidence and they consistently point in one direction: the adolescent brain is particularly vulnerable to the cognitive-load profile that frontier AI tools, like other intensively-engaging digital systems, present.

3.4 The international policy and competency landscape

The international policy environment has converged remarkably quickly. Section 7 develops this in detail; here I note only that as of mid-2026, at least seven major competency or policy frameworks now operate, each constraining or guiding institutional AI deployment in education. These include the EU AI Act (2024); UNESCO's AI competency frameworks for students and teachers (2024); the EU-OECD AILit Framework for primary and secondary education (2025), feeding into the PISA 2029 MAIL assessment (OECD, 2026); the US Department of Labor AI Literacy Framework; Digital Promise's Understand-Use-Evaluate framework; the Digital Education Council AI Literacy framework; and emerging national frameworks in Singapore (Chew, Soon, & Kaur, 2026), the United States (California Department of Education, 2025), and elsewhere. The convergence is not coincidental. Across frameworks, four domains recur with high consistency: critical engagement, creation and collaboration, management and ethics, and design and systems thinking.

The 4D framework offered in this paper is structurally compatible with these convergent international frameworks. Where the international frameworks define the competency outcomes a learner should achieve, the 4D framework defines the cognitive routines a learner must develop to achieve those competencies. The two operate at different levels of pedagogical specification. Table 4 in Section 7 maps the relationship between the 4D framework and the AILit four domains explicitly.

4. The 4D Framework · Delegation, Description, Discernment, Diligence

I propose, and use in my own teaching practice across IFSA Prague, CEA CAPA, and Anglo-American University Prague, a four-stage cognitive framework for human-AI interaction. The four stages are intended to be developmentally and pedagogically sequential, mapping to distinct cognitive routines that must be deliberately taught, in order, with appropriate scaffolding, across the K-12 and tertiary curriculum. The framework is not a competency taxonomy in the sense that the AILit, UNESCO, or Digital Promise frameworks are. It is a cognitive routine framework. It specifies what a learner must learn to do, in the sense of executing a specific cognitive operation reliably, at each stage of development. It is intended to operate beneath competency frameworks as the cognitive substrate from which those competencies are built.

The four stages, briefly: Delegation is the conscious choice of which cognitive tasks to assign to a tool versus to retain for oneself, made with awareness of the reasoning behind the choice. Description is the capacity to specify a problem, context, and desired output to a system with sufficient precision that the system's response is actually useful, and to recognise when the description itself has failed. Discernment is the capacity to evaluate AI output critically, identifying errors, hallucinations, missing context, and inappropriate confidence, without slipping into either credulous acceptance or reflexive rejection. Diligence is the responsible follow-through that closes the loop, verification, attribution, ethical reflection, accountability for outcomes. Each stage builds on the previous. Each can be taught, scaffolded, and assessed. Each requires distinct pedagogical approaches at different developmental tiers.

4.1 Delegation

Delegation is the conscious choice of which cognitive tasks to assign to a tool versus to retain for oneself, made with awareness of the reasoning behind the choice. It is the first cognitive operation in any human-AI interaction, executed every time a learner reaches for the tool, and it is the operation most prone to unreflective default. The pedagogically

interesting question is not whether students delegate. All cognitively functional humans delegate constantly. The question is whether they delegate deliberately, with conscious choice and awareness of the consequences for their own learning, or whether the delegation is unconscious, default, and substitutive.

The empirical work cited in Section 3.1 suggests that the default mode for unguided learners is unreflective over-delegation. Kosmyrna et al. (2025) and Fan et al. (2025) both document patterns in which learners hand the system not just mechanical tasks (formatting, copy editing, boilerplate generation) but the very cognitive operations the assignment was designed to develop in them (premise framing, argument construction, evidence selection). Teaching Delegation, therefore, is teaching the explicit cognitive routine of pausing before reaching for the tool, articulating to oneself what cognitive operation the task is intended to develop, and choosing on that basis whether to delegate it. This is a metacognitive skill. It must be taught explicitly. It will not emerge spontaneously from exposure to the tool.

4.2 Description

Description is the capacity to specify a problem, context, and desired output to a system with sufficient precision that the system's response is actually useful, and to recognise when the description itself has failed. In the popular form this is called prompting, but the educational stake is broader and older. The discipline of describing a problem (to a teacher, to a peer, to oneself, to a machine) is a foundational cognitive skill. It maps onto the philosophical-educational tradition of articulation, the rhetorical tradition of audience analysis, and the practical tradition of specification in any technical discipline. GenAI demands it in a particularly explicit form. A learner who cannot describe what they want will get back fluent nonsense; a learner who can describe will get back something useful. The key pedagogical insight is that this is not a new skill. It is the explicit form of the implicit articulation skill all good education has always cultivated.

Description is also the stage at which the limits of one's own understanding become palpable. A student who cannot articulate to a model what they actually want to know discovers, often for the first time, that the underlying cognitive operation (knowing what one wants to know) is incomplete. This diagnostic function of Description, where the failure of description reveals the prior failure of articulation, makes it one of the most pedagogically valuable stages of the framework. The educator's role at this stage is to surface the diagnostic moment rather than to bypass it.

4.3 Discernment

Discernment is the capacity to evaluate AI output critically. This is where AI fluency intersects with classical critical thinking, epistemological literacy, and source evaluation. A discerning learner reads AI output the way a competent researcher reads a secondary source: looking for evidence, checking citations, noting confident assertions about contested matters, identifying what the source is unlikely to know. The specific challenge GenAI raises here is the surface-confidence problem. Frontier models produce fluent, syntactically polished output that signals confidence regardless of underlying epistemic status. The discernment routines that worked for evaluating, say, a Wikipedia article, where uncertainty is often signalled by sentence structure, citation density, and tone, do not transfer cleanly to AI output, where fluency is uniform.

Discernment is intimately tied to domain knowledge. A learner can only discern in domains where they have enough substantive knowledge to evaluate. This produces a pedagogical paradox. To use AI well, students need domain knowledge. To acquire

domain knowledge in a school system increasingly mediated by AI, students need to be protected from premature AI substitution for the very routines that build domain knowledge. This is one of the strongest arguments for asymmetric AI exposure across developmental tiers (Section 6), and against uniform permission policies that treat AI access as developmentally neutral.

4.4 Diligence

Diligence is the responsible follow-through that closes the loop. Verification of factual claims. Attribution where attribution is owed. Disclosure where institutional norms or honesty require it. Reflection on what was learned, what was substituted, what was bypassed. Ethical accountability for the outcomes of the AI-assisted process. Diligence is what distinguishes a competent operator from a fluent user. It is the stage at which AI fluency intersects with academic integrity, professional ethics, and, in regulated industries, legal accountability. It is, in my view, the stage most consistently neglected in current educational practice, both because it is the least exciting (verification, disclosure, attribution are administratively dull operations) and because it imposes the highest cognitive cost.

Taken together, the four stages provide a developmental sequence: a learner who can Delegate consciously, Describe precisely, Discern critically, and follow through with Diligence is operating at what I term AI fluency. Crucially, fluency is not equivalent to permission. A learner can be permitted to use AI without being fluent (the default at most tertiary institutions today). A learner can be denied permission to use AI without being prevented from developing fluency (the gap between formal policy and shadow curriculum). And a learner can be developmentally too young to develop full fluency in any of the four stages, in which case the pedagogical task is to build the precursor cognitive routines that fluency will eventually be built upon. This is what the framework asks educational institutions to take seriously. Fluency is not a permission decision (Should we allow AI? Yes/No). It is not a tool selection decision (Which AI? GPT/Claude/Gemini). It is not a detection decision (How do we catch AI use? AI-detector/oral exam/in-class). It is a cognitive curriculum decision (What cognitive routines should learners develop, in what order, at what age, under what supervision, with what accountability). It is the developmental sequencing of these four stages, not the binary, that the framework asks institutions to take seriously.

5. Developmental Considerations · The Neuroscience of AI Exposure

The case for asymmetric AI exposure across developmental tiers rests on three converging bodies of evidence: the developmental neuroscience of executive function maturation, the literature on cognitive load and self-regulation in adolescence, and the emerging evidence on intensive digital exposure effects in children and adolescents. None of these literatures was developed specifically with GenAI in mind. Taken together, however, they establish a sufficient *prima facie* case to constrain the default institutional posture.

The prefrontal cortex (PFC), seat of executive function, judgement, and inhibitory control, does not reach mature myelination until the mid-twenties (Diamond, 2013). Working memory capacity, the cognitive resource that supports complex reasoning, sustained attention, and the integration of new information into existing schemata, is structurally limited in childhood and adolescence and improves only with deliberate practice on cognitively demanding tasks. Metacognition, the ability to monitor and regulate one's own

cognitive processes, develops slowly across the school years and is particularly vulnerable to environmental conditions that substitute external regulation for internal regulation. Each of these capacities is precisely what fluent GenAI use requires. Each of these capacities is also what GenAI use can most easily substitute for, in the way an exoskeleton substitutes for muscle development.

The risk is therefore not abstract. There is no plausible developmental mechanism by which a child whose first encounter with a complex cognitive task is an AI-generated solution to it will form the same cognitive routines as a child who struggles with the task first and encounters the AI later, as a verification or extension tool. The first child gets the answer. The second child builds the muscle that produces answers. The first child's prefrontal cortex has no developmental reason to lay down the neural circuitry that sustained struggle would have built. This is the substitution risk, and it is most acute precisely in the developmental window during which executive function is most plastic.

Empirical findings on digital exposure effects in this window are consistent enough to warrant precaution even where direct GenAI evidence is unavailable. Hutton et al. (2022) report associations between screen-based media use and reduced white matter integrity in preschool-aged children, with measurable effects on tracts supporting language and emergent literacy. Lewin et al. (2025) report longitudinal effects of screen media activities on brain function, architecture, and mental health in early adolescence, reporting an average 0.82mm reduction in PFC thickness in high-screen-exposure children versus matched controls, a difference associated with measurable executive function decrements. León-Domínguez (2024) reviews potential cognitive risks of generative transformer-based AI chatbots on higher-order executive functions, noting in particular the risk of substitutive use during developmental periods of high neural plasticity. Gascón-Cánovas et al. (2025) report differential executive function performance in adolescents with problematic smartphone use, with effects detectable on parent-report measures even where controlled lab tasks may underestimate them. Hinze et al. (2025), in a longitudinal study of 8,072 UK adolescents, demonstrate stable bidirectional associations between executive function development and mental health outcomes across the 11–15 age range.

Reggente et al. (2026), in a 104-study scoping review (Pediatric Reports) covering 2022–2025 research on smartphone use and adolescent brain function, document reduced functional connectivity in cortico-subcortical circuits, reduced grey matter volume in the superior and inferior frontal gyrus, and aberrant cortical activation in frontoparietal regions during cognitive conflict tasks among adolescents with high smartphone use. Five recurring themes emerge across the review: psychological disturbances (n = 52 studies), sleep (n = 43), socioemotional function (n = 23), executive function (n = 14), and sensory processing (n = 1). The methodological caveats from Section 3.3 apply: these are correlational findings on intensive digital exposure generally, not on GenAI specifically. They are however the best available proxy evidence, and they consistently point in the direction that the adolescent brain is particularly vulnerable to the cognitive-load profile that intensively engaging digital systems, of which frontier AI is now the most cognitively engaging, present.

The implication is not that screens, digital media, or AI tools cause harm in any simple sense. The implication is that cognitive load, self-regulation demand, and reward structure are non-trivial variables, and that the developing brain is more vulnerable to their mis-calibration than the adult brain. Generative AI introduces all three variables in an unusually concentrated form: it offloads cognitive load instantly, it requires

self-regulation to use constructively rather than substitutively, and its reward structure (immediate, fluent, validating output) is precisely the profile that compromises inhibitory control in adolescent populations. Each of these properties is well-documented as a developmental risk factor independently. GenAI combines them.

From these neurodevelopmental realities, three pedagogical principles follow.

(i) Asymmetric exposure. The intensity, autonomy, and breadth of GenAI integration should scale with cognitive-developmental readiness, not with grade level alone. A 12-year-old and an 18-year-old should not be operating under the same AI access policy. The policy should be calibrated to executive function, working memory capacity, and metacognitive maturity. Where measurement is impractical (as it generally is at institutional scale), reasonable proxies (chronological age, developmental tier) are defensible, but the policy must actually differ across them.

(ii) Substitution risk before scaffolding maturity. Before learners have internalised the cognitive routines a tool is intended to augment, the tool risks substituting for the routine rather than augmenting it. The pedagogical sequencing must therefore introduce AI tools after, not before, the underlying skill is developmentally consolidated. This is the strongest argument for protective postures in primary education and for scaffolded postures in secondary education.

(iii) Friction is pedagogically valuable. The aspect of GenAI that makes it commercially compelling (frictionless answer generation) is the aspect that makes it pedagogically risky. Effective educational integration deliberately re-introduces friction (verification, reflection, disclosure, debate against the model) at every stage. Constructionist pedagogy (Lin, Dai, & Ng, 2025) provides one well-researched mechanism: requiring students to construct artefacts with AI rather than receive outputs from AI. Project-based learning (Thomas, 2000) and computational thinking traditions (Wing, 2006) offer complementary mechanisms.

6. Per-Tier Pedagogical Approach

I propose three structurally distinct integration models across the K-12 plus tertiary spectrum. They are not gradual gradients of the same approach. They differ in kind, calibrated to what each tier of learner can do, what they should do, and what should be protected. Table 1 summarises the proposed posture across the three tiers.

Table 1. Developmentally-aligned posture across educational tiers.

Tier	Age	AI Posture	Primary Pedagogical Focus
Primary	5-11	AI-adjacent literacy. No direct LLM access in classroom.	Foundational reading, writing, arithmetic, attention, social cognition. Critical thinking modelled by teacher.
Secondary	12-17	Scaffolded AI literacy with structured exposure under educator supervision.	All four 4D stages explicitly taught. Discernment and Diligence load-bearing.
Tertiary	18+	Operational AI fluency. Domain-specific integration.	Professional-grade prompting, model selection, agentic systems, ethical deployment, disclosure norms.

6.1 Primary Education (Ages 5-11) · AI-Adjacent Literacy

The defensible posture in primary education is protection of foundational cognition. Children in this developmental window are still consolidating the core skills (reading fluency, arithmetic automaticity, written self-expression, sustained attention, peer cooperation), that all subsequent cognitive work depends on. These skills are built through productive struggle. Substitutive cognitive offloading at this stage is not a pedagogical risk it is a pedagogical failure. A child whose first experience of generating written text is generating it with an AI assistant has not learned to write. They have learned to commission. The neural circuitry of writing is laid down through the act of writing, not through the act of reviewing AI-generated text.

The framework I recommend at this tier is what I call AI-adjacent literacy. Children learn about AI, what it is, what it can and cannot do, why it sometimes sounds confident and is wrong, that it is made by people and reflects choices people have made, without interacting directly with frontier models in the classroom. Concrete pedagogical practices at this tier include teacher-modelled AI interaction in age-appropriate domains (the teacher demonstrates, narrates the reasoning, surfaces what they are choosing to retain versus delegate), simplified algorithmic-thinking exercises (decomposition, pattern recognition, abstraction in the Wing, 2006 sense), introduction to the social and ethical dimensions of automated decision-making in age-appropriate forms, and conversations about how the child distinguishes between machine-generated and human-generated content.

Constructionist approaches developed in pre-GenAI K-12 AI literacy research (Lin, Dai, & Ng, 2025) translate cleanly to this tier. Where the older constructionist tradition had children build robots and simple ML classifiers to learn about AI, the GenAI-era version extends this to having children build small simulations and games that include AI characters whose behaviour the child specifies and interrogates. Su, Ng, and Chu (2025), in a systematic review of 47 empirical studies on K-12 AI literacy, identify constructionism, project-based learning, and embodied cognition as the three approaches with the strongest evidentiary support for foundational AI literacy at this tier. Crucially, none of these approaches requires children to interact directly with frontier consumer LLMs.

This posture is not technophobic. It is developmentally appropriate sequencing. Children at this tier learn to read by reading books that are calibrated to their cognitive stage, not by being handed Shakespeare. The principle is the same. Direct GenAI exposure may have a legitimate role at the upper bound of this tier (ages 9-11) for specific, supervised, scaffolded uses, particularly in service of teaching Description and the first stages of Discernment. But the default posture should be protection, and the burden of justification should fall on those advocating direct unsupervised use, not on those protecting the developmental window.

6.2 Secondary Education (Ages 12-17) · Scaffolded AI Literacy

Secondary education is the developmentally critical window. Students at this tier are cognitively capable of using frontier AI tools (working memory, executive function, and metacognitive capacities are sufficiently developed) but their cognitive routines are still consolidating. The adolescent brain is also at peak vulnerability to the reward-structure properties that compromise inhibitory control (Steinberg, 2008; Reggente et al., 2026). This is the tier at which the framework's full weight must be applied. Get this stage wrong and the institutional cost is visible across the credential system. Get it right and the resulting tertiary cohorts arrive with the cognitive foundations operational AI fluency requires.

All four 4D stages should be taught explicitly across the secondary curriculum, with progressive sophistication. Delegation work begins with metacognitive exercises: the student articulates, before each AI interaction, what cognitive operation the task is intended to develop in them, and chooses on that basis whether to delegate it. Description work emphasises iterative refinement: students learn to evaluate whether a model response actually addressed their question, identify which features of the description failed, and revise. Discernment is the most pedagogically intensive stage at this tier. Students learn to verify AI outputs against authoritative sources, identify hallucinations and inappropriate confidence, recognise the limits of model knowledge (training cutoff, training-data bias, domain-specific weakness), and disagree productively with the model. Diligence is taught through explicit attribution norms, disclosure requirements, and reflection on what was substituted versus what was built.

The pedagogical structure that supports this is what I call scaffolded AI literacy. AI tools are introduced into the curriculum deliberately, sequenced with the underlying skills they are intended to augment, with explicit teacher guidance on the cognitive operation the student should be performing and the operation the AI should be performing. The classroom design that supports this includes structured AI interactions (clearly bounded tasks, time, and goal), educator review of AI interactions (sampled, not surveilled), and explicit instruction on prompting, verification, and disclosure as transferable skills.

Two further pedagogical commitments distinguish this tier. First, assessment redesign: traditional take-home and unsupervised assessments no longer reliably measure the cognitive operations they were designed to measure. Assessment must move toward process-oriented evidence (learning logs, version history, think-aloud protocols), performance-based tasks (collaborative problem-solving, design challenges), and in-class supervised demonstration of routines that have been built up across the curriculum (ATC21S framework, in OECD 2030 Learning Compass). Second, transparent disclosure norms: students learn, from the first introduction of AI tools, that disclosure of AI use is expected, normal, and itself part of the assessment, rather than a confession to be concealed. The disclosure framing transforms academic integrity work from policing to pedagogy.

Roe, Furze, and Perkins (2026), in their digital plastic metaphor framework, propose that AI literacy in this tier should be explicitly framed as critical AI literacy: students learn to evaluate AI not only as a tool but as a sociocultural artefact embedded in training data, business model, and political-economic context. This framing pairs naturally with the 4D framework. Discernment becomes not only output-evaluation but also system-evaluation. Diligence includes ethical reflection on the AI's existence, not only on the specific outputs the student has used.

6.3 Tertiary and Higher Education (18+) · Operational AI Fluency

Tertiary learners are cognitively positioned for full AI fluency. The pedagogical task at this tier is operational mastery within a domain. Students should leave their degree programme able to operate frontier AI tools at professional standard within their discipline, including domain-specific prompting, model selection, multi-step task decomposition for agentic systems, and ethical deployment within professional norms. Tertiary curricula should integrate AI use as a graduated, assessed competence in every discipline where it is professionally relevant, which by 2026 is essentially every discipline.

Concrete operational components include: domain-specific prompt engineering (the conventions, sources, and evaluation criteria specific to the discipline), model-selection literacy (when to use which model, what the trade-offs are, how to evaluate suitability),

agentic and multi-step workflow design (how to compose tool use across multiple invocations, how to build verification into the workflow), professional ethics and disclosure norms (discipline-specific, calibrated to professional bodies and emerging regulation), and the operational implications of regulation including the EU AI Act for graduates entering AI-deploying industries (European Union, 2024).

The empirical evidence on tertiary AI integration supports cautious optimism about learning outcomes when integration is structured. Cunningham et al. (2025), in their year-long Tennessee study, report substantial positive effects on graduate-student motivation and skill acquisition when ChatGPT integration is paired with explicit instructional framing. Raptopoulou (2025) reports that ChatGPT supports academic literacy development when integrated with clear pedagogical goals. The meta-analysis by Deng et al. (2026) reports moderate positive effects on learning outcomes across 35 studies. The qualifier in each case is the same: positive effects depend on structured, scaffolded integration, not on permissive default. Unstructured access produces the metacognitive laziness Fan et al. (2025) document; structured integration produces the operational fluency Cunningham et al. (2025) report.

The pedagogical sequencing across tertiary education should therefore be: explicit instruction on the 4D framework in the first year (if not already mastered from secondary education); domain-specific integration across the middle years (the 4D framework applied within the discipline's epistemic, methodological, and ethical conventions); and capstone or final-year work that explicitly assesses operational fluency through realistic professional tasks. The model is closer to a clinical placement or studio review than to a traditional written examination. AI fluency is operationalised competence, demonstrated under realistic conditions, with the same professional ethics and accountability the student will require, including in regulated domains where the standard is high.

6.4 Sample Activities and Operational Examples by Tier

Abstract framework specifications are necessary but insufficient for institutional implementation. This sub-section offers concrete examples of activities at each tier that operationalise the 4D framework in classroom practice. The examples are drawn from my own teaching practice, the literature reviewed in Section 3, and the AILit framework's classroom scenarios (European Commission & OECD, 2025). They are not prescriptive. They are intended to illustrate the kind of pedagogical design the framework calls for, against which institutional curriculum work can be tested and refined. Table 3 summarises a representative activity set.

Table 3. Representative classroom activities operationalising the 4D framework.

Tier	4D Stage	Sample Activity	Pedagogical Aim
Primary (9–11)	Discernment	Teacher-led inspection of an AI response showing a confident factual error. Children identify the error using a printed primary source.	Establish that confident-sounding output is not evidence of correctness.
Primary (9–11)	Delegation (precursor)	Class discussion: should we ask the AI for help with our spelling, or should we work it out ourselves first? When does each make sense?	Surface conscious choice as a learnable skill before any tool use.
Secondary (12–14)	Description	Students attempt to get an AI tool to generate a useful answer to a research question, with three iterations of prompt refinement, documented and discussed.	Articulation as a transferable cognitive skill.

Tier	4D Stage	Sample Activity	Pedagogical Aim
Secondary (15–17)	Discernment	AI-generated source list verification exercise. Each student receives an AI-generated bibliography and must verify which citations are real and which are hallucinations.	Build operational verification routine.
Secondary (15–17)	Diligence	Process portfolio: every AI-assisted assignment includes a structured disclosure log specifying what was delegated, why, and what was verified.	Habituate disclosure and reflection as normal practice.
Tertiary	Delegation	Discipline-specific decision matrix: students develop their own criteria for when to use which model, with which workflow, for which task in their field.	Professional-grade delegation judgement.
Tertiary	Diligence	Regulated-industry simulation: students must produce work to a quality standard with full disclosure, attribution, and verification trail compliant with EU AI Act Article 4 expectations.	Bridge to professional accountability.

Several design principles cut across the examples in Table 3. First, the cognitive routine is named explicitly. Students are told what they are learning to do (delegation, description, discernment, diligence), not only what task they are completing. The metacognitive framing matters; it converts a tool-use activity into a routine-formation activity. Second, the routine precedes the tool. In several of the activities (notably Description and Delegation work), students articulate the cognitive operation before the AI is involved. The AI then becomes the test environment for an already-formed cognitive routine, not the substitute for forming it. Third, friction is deliberately preserved. Verification, disclosure, and iteration are not bypassed because the tool allows them to be; they are required because the curriculum requires them. The friction is the lesson.

Fourth, assessment is embedded. Each activity produces evidence (the prompt iterations, the verified bibliography, the disclosure log, the decision matrix) that can be assessed for the cognitive operations the activity targets. This addresses the assessment-validity collapse identified in Section 2: the assessment evidence is the routine-formation evidence, not the AI-substitutable output. Fifth, the activities scale across the curriculum. Secondary-school activities prepare for tertiary activities. Primary-school precursors prepare for secondary activities. The framework is intended to be implemented across the K-16 spectrum as a developmentally coherent sequence, not as discrete grade-specific units.

Two operational pitfalls deserve explicit attention. The first is what I term token-disclosure: the formal addition of an 'I used AI for X' line at the end of student work, with no engagement with what was delegated, why, or what was verified. Token disclosure satisfies surface compliance without the underlying Diligence routine the disclosure was designed to surface. Effective Diligence pedagogy requires structured disclosure formats (what was delegated, why, what was verified, what was left unverified, what the student would do differently next time) and educator engagement with the disclosure content. The second pitfall is theatre-of-friction: the addition of formal friction (verification requirements, version-history requirements) that students perform mechanically without cognitive engagement. Theatre-of-friction produces the appearance of routine without the substance. The remedy is the same as for token-disclosure:

educator engagement with the substance of the routine evidence, not just its formal presence.

7. The International Policy Landscape · Convergent Frameworks

The international policy environment around AI in education has converged with unusual speed across 2023–2026. As of mid-2026 at least seven major frameworks now shape institutional decision-making, with substantial conceptual overlap and growing operational alignment. This section summarises the principal frameworks and maps their relationship to the 4D framework proposed in this paper. The convergence represents an opportunity: a globally interpretable competency vocabulary is forming, against which national systems can be evaluated and between which institutional reform can be coordinated. It also represents a constraint. Institutions deploying GenAI without engagement with this emerging policy infrastructure are increasingly out of step with both regulatory expectation and professional norm.

7.1 The EU AI Act and Article 4 AI literacy obligation

The European Union's AI Act (Regulation (EU) 2024/1689) entered into force on 1 August 2024 and is being phased in through 2027 (European Union, 2024). Article 4 imposes a binding AI literacy obligation on providers and deployers of AI systems: they must ensure that their staff, and anyone using the systems on their behalf, have an adequate level of AI literacy. For educational institutions this creates a twofold obligation. First, institutions deploying AI tools, whether for administrative, teaching, or assessment purposes, are themselves deployers within the meaning of the Act, with a direct compliance obligation on staff AI literacy. Second, institutions whose graduates will enter regulated industries are training the staff to whom the Article 4 obligation will apply at their future workplace. The downstream professional obligation makes AI literacy not optional curriculum content but professionally required preparation.

7.2 UNESCO frameworks for students and teachers

UNESCO released two AI competency frameworks at Digital Learning Week in September 2024, one for students and one for teachers. The teacher framework (Miao & Cukurova, 2024) defines 15 teacher competencies across five dimensions: human-centred mindset; ethics of AI; AI foundations and applications; AI pedagogy; and AI for professional development. The student framework defines parallel competencies at the learner level. UNESCO's framing emphasises ethical and human-centred deployment, particularly in contexts where the technology, infrastructure, or institutional capacity to ensure responsible use are uneven. Global readiness against the UNESCO frameworks has been estimated at under 4% (Welker, 2025).

7.3 The EU-OECD AILit Framework

The European Commission and OECD jointly released Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education (the AILit Framework) in May 2025 as a review draft, with the final version scheduled for 2026 (European Commission & OECD, 2025). The framework organises AI literacy around four domains, with 22 competences distributed across them: Engaging with AI (understanding when and how AI is present, critically evaluating outputs); Creating with AI (collaborating with AI tools to support problem-solving and creativity, with attention to ethical implications); Managing AI (responsible decision-making about personal data, agency, and societal impact); and Designing AI (foundational understanding of how AI systems are

designed and built). The framework was developed with input from international experts across multiple countries and was informed by consultations with nearly 1,000 educators, researchers, policymakers, and stakeholders.

7.4 PISA 2029 MAIL Assessment

The OECD released the first draft of the PISA 2029 Media and Artificial Intelligence Literacy (MAIL) assessment framework in February 2026 (OECD, 2026), titled *Navigating an Evolving Digital World*. The assessment will be administered as part of the PISA 2029 cycle, with results expected in December 2031. The MAIL assessment will test 15-year-old students' ability to engage critically and responsibly in environments shaped by media platforms and AI. The institutional implication is significant. AI literacy will, by 2029, be a measured, internationally comparable competence. National educational systems will be benchmarked against each other on it. This is the strongest policy signal yet that AI literacy is moving from optional enhancement to required curriculum.

7.5 National and sector frameworks

Beyond the EU and OECD work, multiple national and sector-specific frameworks now operate. The US Department of Labor has released an AI Literacy Framework with five foundational content areas and seven delivery principles, intended to guide workforce-level AI literacy efforts across industries (AI Literacy Institute, 2026). Digital Promise has proposed an AI literacy framework organised around three domains, Understand, Use, and Evaluate, intended for K-12 and educator use (Digital Promise, 2024). The Digital Education Council has released an AI literacy framework emphasising human-centred competencies (critical thinking, creativity, emotional intelligence). The California Department of Education has released AI literacy guidance organised around developmentally appropriate progressions across grade levels (California Department of Education, 2025). Singapore's Institute of Policy Studies has released a working paper proposing a unified national AI literacy framework synthesising international approaches with Singapore-specific implementation (Chew, Soon, & Kaur, 2026).

Across these frameworks, four substantive domains recur with high consistency: critical engagement with AI (corresponding broadly to Discernment in the 4D framework), creation and collaboration with AI (drawing on Delegation and Description), management and ethical deployment of AI (Diligence), and design or systems-level understanding of AI (a broader competency the 4D framework does not address directly, instead leaving it to disciplinary integration at the tertiary tier).

7.6 The 4D framework in relation to convergent competency frameworks

Table 2 maps the 4D framework's cognitive routines to the four AILit domains and to the four-domain consensus that has emerged from across the international frameworks. The framework offered here operates at a different level of pedagogical specification. Where the competency frameworks specify the outcomes a learner should be able to demonstrate, the 4D framework specifies the cognitive routines a learner must develop in order to demonstrate those outcomes. The two are complementary, not in competition. An institution that adopts the AILit framework as its competency outcome specification, and the 4D framework as its cognitive routine specification, has a complete pedagogical architecture: what students should be able to do, and what cognitive routines they need to build to be able to do it.

Table 2. Mapping the 4D framework to the AILit four domains.

AI/IT Domain	Domain Description	4D Routines Engaged	Tier Emphasis
Engaging with AI	Critical evaluation of AI outputs and recognition of AI presence in tools.	Discernment (primary); Delegation (secondary)	All tiers, depth increases
Creating with AI	Collaboration with AI for problem-solving and creativity, with ethical attention.	Description (primary); Delegation, Diligence	Secondary onward
Managing AI	Responsible decisions about personal data, agency, societal impact.	Diligence (primary); Discernment	Secondary and tertiary
Designing AI	Foundational understanding of how AI systems are designed and built.	Discernment extended to systems level	Tertiary, optional in upper secondary

8. Tool Selection · Why Model Architecture Matters in Education

Educational institutions deploying GenAI to learners, particularly to minors, make a non-trivial choice when they select a frontier model. The choice is often treated as commodity selection (which subscription, which features, which integration), but it is, in my view, a pedagogical choice with consequences proportionate to the developmental vulnerability of the learner population. This section examines what an institutional choice looks like when made on pedagogical and safety criteria rather than feature or cost criteria.

Frontier large language model development has bifurcated, over 2022–2026, between systems whose alignment methodology is publicly documented in substantive technical detail and systems whose alignment methodology is undisclosed or only partially disclosed. This is not a marketing distinction. It is a methodological one. Anthropic, the developer of Claude, has published detailed technical papers on its Constitutional AI methodology (Bai et al., 2022), in which the model is trained against an explicit, public set of constitutional principles (derived from sources including the UN Universal Declaration of Human Rights), and the post-training reinforcement-learning-from-AI-feedback process is documented in sufficient detail for external scrutiny. Anthropic also publishes a Responsible Scaling Policy (Anthropic, 2026) documenting the safety thresholds and mitigations applied as model capabilities scale. OpenAI publishes model and system cards. Google DeepMind publishes technical reports of varying detail. The level of methodological transparency, however, differs materially across vendors.

For deployment to adult professional users, this transparency gradient may be a matter of professional or research preference. For deployment to minors in educational settings, it is a fiduciary issue. Educational institutions stand in loco parentis for many of the children and adolescents to whom they deploy these systems. The institution is making a judgement, on the family's behalf, about which cognitive partner the child will spend large amounts of cognitively formative time with. That judgement should be defensible against the question: what evidence does the institution have, that the system in question is aligned to produce outputs consistent with the institution's values, the family's values, and the child's developmental safety. Methodology transparency makes that defensibility evaluable. Methodology opacity makes it a matter of vendor trust alone.

Anthropic's Constitutional AI methodology (Bai et al., 2022) is the most extensively documented alignment methodology currently available in production frontier systems. The methodology, briefly: rather than relying solely on human feedback to shape model behaviour, the model is trained against an explicit set of constitutional principles, and a second model (also Anthropic-trained) provides feedback on compliance with those principles, with the constitution and the training process documented publicly. This produces a system whose alignment to a particular value structure is auditable, and whose value structure is itself contestable. The Responsible Scaling Policy (Anthropic, 2026) extends this methodological transparency to the scaling-decision process. The RSP framework has not been without controversy. TIME reported in February 2026 that Anthropic relaxed one of the original RSP's central pre-training commitments, which generated substantive criticism (Perrigo, 2026). But the broader public documentation of methodology and methodology revisions is itself the asset.

I am not arguing that Claude is necessarily the best frontier model for any given educational task on capability or cost grounds. Capability comparisons are domain-specific and rapidly shifting. I am arguing that methodology transparency is a defensible criterion for educational institutional selection, that it is reasonably operationalised through publicly documented alignment research and responsible scaling commitments, and that the institutional default should weight it heavily in deployment decisions affecting minors. A more capable model with opaque methodology is not categorically preferable to a slightly less capable model with transparent methodology, for educational deployment to children.

Three operational criteria follow. (i) Methodology transparency. The vendor publishes technical detail of alignment methodology sufficient for external evaluation. (ii) Regulatory cooperation. The vendor cooperates with relevant regulatory frameworks (EU AI Act compliance, national child-protection regulation, sector-specific frameworks where applicable). (iii) Safety research investment. The vendor demonstrates material investment in safety research, with publications and operational practices that suggest the safety frame is not purely marketing. These three criteria are not exhaustive, and they will become more refinable as the field matures, but they offer institutions a defensible starting position for the model-selection conversation.

9. Safety, Ethics, and the Protection-by-Design Principle

Beyond model selection, institutional deployment of GenAI raises a cluster of safety and ethics considerations that current educational policy addresses unevenly. I group them under five headings, with reference to relevant regulatory and research contexts.

(i) Data protection and minor consent. Frontier AI deployment to learners under 18 implicates the EU General Data Protection Regulation, national child-protection frameworks, and emerging EU AI Act provisions. Many current institutional deployments do not satisfy these obligations cleanly. Institutions should document the legal basis for processing, the data retention regime, and the consent architecture for any AI deployment to minors. The default should be data-minimisation: AI tools used in educational settings should be configured to retain the minimum personal data necessary for the educational purpose, with retention limits aligned to educational rather than commercial timescales. Where vendor architectures default to broader retention, the institution carries the obligation to negotiate or to select an alternative.

(ii) Content exposure and age-appropriate guardrails. Frontier models can, in principle, generate or discuss content inappropriate for younger learners. Institutional deployments

should include age-banded guardrails, content filters configured to the developmental tier, and clear escalation paths for inappropriate exposure. Educators should be trained to recognise the signs that a student may have encountered inappropriate content via AI tools, and the school should have a documented response protocol. Critically, the institution carries the obligation to do this proactively, not reactively, the first time inappropriate content reaches a student is too late.

(iii) Cognitive over-reliance and addiction patterns. Recent work has identified what some authors describe as Generative AI Addiction Syndrome (Kooli et al., 2025; Gerlich, 2025), characterised by compulsive use patterns, diminished tolerance for cognitive effort, and withdrawal-like discomfort when the tool is unavailable. Institutional policy should treat usage-frequency limits as protective, not puritanical. This is structurally similar to the institutional response to problematic gaming or social media use, which the literature has documented extensively (Reggente et al., 2026): the school cannot prevent use entirely, but it can structure the school environment such that AI use is not the default for cognitive tasks, and it can teach students to recognise and self-regulate problematic use patterns.

(iv) Bias, fairness, and the representational politics of training data. Frontier models embed the demographic, linguistic, and epistemic biases of their training corpora. Educational deployment should include explicit pedagogical attention to this fact. Students should learn to interrogate AI outputs for bias, not be passively exposed to outputs that smuggle bias under fluency. This pedagogical obligation is particularly acute for learners from groups underrepresented in training data, whose own languages, cultures, and ways of knowing are statistically underweighted in model outputs. A learner who internalises model fluency as authoritative is internalising the training corpus's representational politics. Critical AI literacy (Roe, Furze, & Perkins, 2026; Illingworth, 2026) provides one well-developed approach to teaching this critical posture.

(v) Environmental cost. Frontier model inference and training carry non-trivial energy and water footprints. Roe, Furze, and Perkins (2025), in their Critical AI Literacy framework, place environmental cost among the four core interrogation domains for AI literacy. Institutional deployment should account for this transparently. Where institutions opt for unlimited AI access policies, the environmental implications should be disclosed and reflected upon as part of the curriculum, not concealed in the licensing arrangement.

Taken together, these considerations argue for what I call protection-by-design: AI deployment whose default configuration is protective, requiring affirmative justification to relax protections rather than affirmative effort to add them. Protection-by-design is the educational application of the privacy-by-design principle now embedded in EU GDPR and emerging in AI regulation. It places the burden of justification on those advocating relaxed protection, not on those advocating protection. For institutions deploying GenAI to minors, this is, in my view, the only defensible default architecture.

10. Educator Adaptation · From Content Deliverer to Cognitive Coach

The role of the educator is shifting. In a world of fluent on-demand information, the value an educator adds is no longer principally the delivery of content. It is the modelling, scaffolding, and assessment of cognitive practices that students cannot acquire from the model alone: discernment, productive disagreement, sustained reasoning under uncertainty, ethical judgement, intellectual courage. The transition is from content

deliverer to cognitive coach. Some educators welcome it. Many do not. The institutional and personal stakes of the transition are high enough that the framework cannot ignore it.

This shift is, in my experience teaching applied AI to American and European students at the tertiary level, both liberating and threatening. It liberates educators from the parts of teaching that AI now performs reasonably well (basic explanation, differentiation of explanation by level, drafting of routine materials). It threatens those whose professional identity is built on those parts. The framework I propose here is, among other things, a roadmap for educator preparation that takes the threat seriously while clarifying what the educator now adds that the AI cannot.

Effective educator preparation in this transition includes, in my view, four elements. First, operator-level AI fluency. Educators must themselves be competent operators of the systems they are teaching students to use. This is a non-negotiable prerequisite. An educator who has never composed a sustained piece of work with an AI assistant cannot effectively teach Discernment or Diligence. Operator fluency is acquired through practice, supported by structured professional development. Anthropic's AI Fluency for Educators course (Anthropic, 2025) offers one structured path. UNESCO's teacher competency framework (Miao & Cukurova, 2024) provides a more comprehensive specification.

Second, pedagogical reframing. Explicit professional development on what it means to teach when students have constant access to a fluent answer machine. This is not technical training in AI tools. It is pedagogical retraining on how to design tasks, sequence instruction, and structure assessment such that the cognitive routines the curriculum is intended to develop are actually developed. The shift from content delivery to cognitive coaching is real, but it is also taught, not assumed.

Third, assessment redesign. Practical capacity to design assessments that remain valid under AI access. Traditional take-home assessment is no longer fit for purpose in most disciplines. Replacement designs include in-class supervised demonstration, oral defence of submitted work, portfolio assembly with version history, process-evidence tracking, and project-based assessment with public artefacts. Each of these has its own design constraints and trade-offs. Educators need explicit training in selecting and operating them.

Fourth, ethical and regulatory literacy. Sufficient grounding in EU AI Act provisions, national child-protection frameworks, and disciplinary professional ethics to operate competently within them. Educators are now, whether they know it or not, deployers of AI systems within the meaning of the EU AI Act. The Article 4 AI literacy obligation falls on them. Institutions that fail to provide this grounding are setting their educators up for compliance failures that are, ultimately, the institution's own.

Across all four elements, the principle is the same: the educator's authority in the AI era derives from cognitive coaching, ethical modelling, and protected attention to the student's developmental trajectory, not from being the source of content. This is, in many ways, a return to older models of pedagogy (the tutor, the Socratic guide, the disciplinary mentor) that the industrial-scale content-delivery model temporarily displaced. The shift is, properly framed, not a threat to the profession but a clarification of what was always its core.

10.1 A Three-Stage Professional Development Pathway

Institutional implementation of the educator preparation requirements above benefits from a staged pathway. Educators do not move from AI-novice to cognitive-coach in a single workshop. The transition is more honestly understood as a multi-year professional development arc, comparable in scope to integrating any major curricular shift (the introduction of computational thinking, the integration of media literacy, the embedding of intercultural competence). Table 4 outlines a three-stage pathway with stage-appropriate professional development components, evaluation indicators, and recommended timeframes. The pathway is not the only possible design but offers an institutionally tractable starting point.

Table 4. Three-stage educator professional development pathway.

Stage	Components	Evaluation Indicators	Timeframe
Stage 1 · Foundational Fluency	Operator-level AI use; basic prompt engineering; tool-selection literacy; introduction to AILit and UNESCO frameworks; EU AI Act Article 4 orientation.	Educator can produce useful output across at least three discipline-relevant task types using a frontier model. Educator can articulate the institution's tool-selection rationale.	3–6 months
Stage 2 · Pedagogical Reframing	Task and assessment redesign workshops; 4D framework integration; cognitive coaching micro-teaching; subject-specific implementation planning.	Educator has redesigned at least one assessment that remains valid under AI access. Educator can model the 4D routines in a classroom observation.	6–12 months
Stage 3 · Institutional Leadership	Curriculum design contribution; peer coaching; participation in institutional AI policy; engagement with external frameworks (PISA 2029 MAIL, national bodies).	Educator leads a curriculum-redesign workstream or peer-coaching cohort. Educator contributes to institutional policy documents on AI integration.	12–24 months

Several implementation principles cut across the three stages. First, the pathway is sequential but not strictly linear; some educators will move through Stage 1 quickly and benefit from accelerated entry to Stage 2, particularly those with existing technology or computational backgrounds. Others will need substantially longer at Stage 1, particularly those in disciplines where the tool's domain-specific utility is less obvious or where prior teaching practice has not engaged with technology mediation. Differentiated pacing is essential. Second, the pathway should be institutionally supported, not individually borne. Time, peer coaching infrastructure, and institutional recognition of progression matter as much as the formal content. Educators who progress through the pathway are providing institutional value comparable to professional certification in other domains; the institutional response should match. Third, the pathway should be visible to external stakeholders, parents, students, regulators, accreditation bodies. The institution's educator-readiness state is increasingly a question external parties will ask, and an institution that can answer with a documented pathway is in materially better position than one that cannot.

An additional consideration is the relationship between institutional and individual educator effort. Some institutions are well positioned to provide the pathway internally, with dedicated educational technology, learning sciences, or pedagogy specialists who can lead the work. Others will need to procure external professional development, partner with universities, or join consortia. Anthropic's AI Fluency for Educators course

(Anthropic, 2025), UNESCO's competency framework training resources (Miao & Cukurova, 2024), and the AILit Framework's classroom scenarios (European Commission & OECD, 2025) provide one set of structured external resources. National AI literacy initiatives in the United States, Singapore, and elsewhere provide complementary infrastructure. The framework is agnostic about which specific external partnerships an institution should use; it is not agnostic about whether such support is needed, which for most institutions it materially is.

11. Forward Look · 2027-2030

Predicting AI capability trajectories at three to five years is unreliable. Predicting institutional responses to them is even more so. I offer here not predictions but the planning assumptions I find most defensible, on which institutional reform decisions in the 2026-2030 horizon can reasonably be based. These assumptions are calibrated cautiously; institutions building on them should expect to revisit them annually as evidence accumulates.

Assumption 1: Continued capability progression. Frontier model capability will continue to advance at rates that exceed institutional curriculum revision cycles. Specific capability frontiers, multi-step reasoning, sustained agentic task execution, long-context coherence, multimodal integration, will continue to expand. Inference costs will continue to fall, making frontier capability accessible to younger and economically disadvantaged learners at rates that outpace institutional ability to extend AI literacy curricula. The institutional implication is that any specific tool training will obsolesce within roughly twelve months. Cognitive routine training (the 4D framework) should be more durable, because the routines required to use AI well, conscious delegation, precise description, critical evaluation, diligent follow-through, are transferable across tool generations.

Assumption 2: Agentic systems become educationally consequential. The transition from single-turn chatbot interactions to multi-step agentic operation (composing tools, executing tasks, returning structured deliverables) is now well underway in commercial deployment and will reach educational deployment by 2027-2028. Educational integration designed for the chatbot paradigm will not transfer cleanly. Agentic systems raise distinct pedagogical issues: how does a student supervise a multi-step process they did not perform? How does the educator assess the cognitive operations performed by the student versus the operations delegated to the agent? The 4D framework is extensible to agentic contexts, but the operational specification of each stage will require adaptation as agentic capability extends.

Assumption 3: Credential erosion accelerates. The gap between what graduates can credibly claim under AI access and what they can perform unaided will widen. Credential value, particularly for credentials primarily assessed through unsupervised written work, will decline correspondingly. Employers and downstream institutions will respond, are already responding, with supplementary verification mechanisms (interview-based competency assessment, project-portfolio review, professional certification, work-sample tasks). The credential-trust problem identified in Section 2 falls hardest on institutions whose credentials are most reliant on the legacy assessment forms. Educational institutions that do not address this directly (through assessment redesign, transparent disclosure norms, and competency-based supplements to traditional credentials) will see the market value of their credentials erode at rates that endanger their economic model.

Assumption 4: Policy and regulatory consolidation. The international policy convergence documented in Section 7 will continue. By 2029, the PISA 2029 MAIL assessment results

will provide the first internationally comparable measurement of AI literacy outcomes (OECD, 2026). Member states whose systems perform poorly will face political pressure to reform. The EU AI Act compliance obligations will be fully phased in by 2027 (European Union, 2024), with Article 4 AI literacy requirements operative across both providers and deployers. Educational institutions in EU jurisdictions will face direct compliance obligations on staff AI literacy. Non-EU systems will be incentivised to maintain comparability for student mobility and qualification recognition purposes. The institutional implication is that AI literacy curricula in the 2027–2030 window are no longer voluntary enhancements; they are increasingly required infrastructure.

Assumption 5: Differential institutional response will produce divergent outcomes. The institutions that begin operational reform now, with deliberate curriculum redesign, educator preparation, assessment redesign, and protection-by-design infrastructure, will arrive at the 2029 PISA measurement point in materially better position than institutions that defer reform. The divergence will compound across the decade. The window for low-cost adaptive reform is narrowing. By 2028 the institutions that have not begun reform will face the compound costs of catching up (teacher retraining, curriculum overhaul, assessment redesign, parental and student trust rebuilding) under increasing external pressure.

I offer four concrete operational recommendations for institutions operating in the 2026–2028 window. First, conduct an AI literacy audit against the AILit framework (European Commission & OECD, 2025) and the UNESCO competency frameworks (UNESCO, 2024; Miao & Cukurova, 2024). Identify the institutional gap explicitly. Second, prioritise educator preparation. Operator fluency for all educators, pedagogical reframing for those teaching cognitively demanding material, assessment redesign training for assessment leads. Third, redesign assessment across the curriculum, beginning with the disciplines most exposed to AI substitution risk. The principle is that assessment should measure the cognitive routines the curriculum is supposed to develop, not the outputs the curriculum is supposed to produce. Fourth, deploy AI tools to learners under protection-by-design defaults, with methodology transparency, regulatory cooperation, and safety research investment as primary selection criteria for adopted tools (Section 8).

11.1 Three Scenarios for 2030

Scenario planning is not prediction; it is a structured way to make institutional decisions robust to a range of plausible futures. I offer three scenarios for the 2030 state of AI integration in education, calibrated to the assumptions above. Each is internally coherent, none is exclusively likely, and the institutional reform decisions of the 2026–2028 window will materially shape which of the three the system actually inhabits.

Scenario A · Convergent reform. Major national systems adopt operationalised versions of the AILit, UNESCO, and emerging national frameworks, with PISA 2029 MAIL data providing the first internationally comparable measurement of progress. EU AI Act compliance is fully phased in by 2027 and operational across deployer institutions by 2029. Educator preparation pathways are institutionally normalised across the OECD and partner countries. The shadow curriculum recedes as institutional curriculum closes the gap with student practice. The credential-trust problem stabilises as assessment redesign matures. This scenario requires sustained political commitment, substantial investment in educator preparation, and continued international policy cooperation. It is achievable but not inevitable.

Scenario B · Two-track divergence. Some national and institutional systems reform aggressively; others defer. The 2029 PISA MAIL data reveal substantial cross-national divergence in AI literacy outcomes. A two-track system emerges in which graduates of reformed institutions enter the labour market with materially higher operational AI fluency than graduates of deferred-reform institutions. Credential value diverges accordingly. International student mobility shifts toward reformed systems. The 2030 picture is one of widening inequality both within and between national systems, with the institutions that began reform in 2026–2028 compounding their advantage. This is, in my assessment, the most likely scenario under current policy trajectories.

Scenario C · Crisis-driven reform. Institutional reform is deferred too long, and the credential-trust problem precipitates a discrete institutional crisis: a major employer-side rejection of credentials, a high-profile assessment fraud scandal, a comprehensive regulatory intervention, or a public-trust collapse of significant magnitude. Reform follows under crisis conditions, with higher costs, less coherence, and worse outcomes than would have followed orderly reform in the 2026–2028 window. Some institutions do not survive the transition. This scenario is not predicted but is not implausible at sub-system level; specific disciplines, institutions, or national systems may inhabit it even where others inhabit Scenarios A or B.

The strategic implication for institutional leadership is straightforward. Reform begun in the 2026–2028 window is robust across all three scenarios. Reform deferred is robust only to Scenario A. Decision-making under genuine uncertainty about which scenario the institution will face should weight robustness. The framework offered in this paper is designed to be implementable across all three scenarios, with the institutional pathway in Sections 6, 7, 8, 9, and 10 forming a coherent reform agenda that retains value regardless of which scenario materialises.

12. Limitations and Future Research Directions

This paper has limitations that should be transparent to readers and that constitute the agenda for the empirical work the framework ultimately requires. I address them in five clusters.

(i) The 4D framework is theoretically motivated but empirically untested. The framework is built on a synthesis of converging literatures (cognitive offloading research, developmental neuroscience, AI literacy frameworks) combined with operational experience in tertiary teaching. It has not been subjected to controlled empirical evaluation, and its specific developmental sequencing claims (asymmetric exposure across tiers; routine before tool) are not directly tested. The framework's testable predictions, for example that students trained explicitly on the 4D routines will demonstrate better Discernment and Diligence outcomes than students given equivalent AI access without the routine training, await empirical examination.

(ii) The neurodevelopmental evidence base concerns digital exposure generally, not GenAI specifically. Section 5 builds the case for asymmetric exposure on proxy evidence from screen, smartphone, and intensive digital exposure research. The proxy is defensible (GenAI shares the relevant cognitive-load and reward-structure properties of the proxy stimuli) but it is a proxy. Direct longitudinal research on GenAI exposure effects on the developing brain does not yet exist in robust form. It is a critical priority for the next decade of educational neuroscience research.

(iii) Cultural and linguistic specificity. The framework is developed primarily in a Western European and North American tertiary-education context, with secondary attention to

Central and Eastern European K-12 systems. Its transferability to non-Western educational contexts, different curricular traditions, and learners whose first languages are systematically underrepresented in frontier model training data, is asserted rather than demonstrated. Validation work in diverse educational contexts is required.

(iv) Tool selection criteria are aspirational rather than operational. Section 8 proposes methodology transparency, regulatory cooperation, and safety research investment as institutional selection criteria. Each of these is somewhat operationalisable but none is yet supported by a robust, vendor-comparable evaluation infrastructure. The work of converting these criteria into a usable institutional evaluation framework is itself a research and policy agenda. The 4D framework's recommendations on tool selection should be read as a starting position, not a finished evaluation methodology.

(v) Pace of change. The empirical evidence base on GenAI in education is moving faster than any single paper can synthesise. Claims and recommendations offered here reflect the state of evidence as of mid-2026 and should be revisited as further empirical work, particularly the PISA 2029 MAIL data (OECD, 2026), the longer-cycle empirical work emerging from the EU's Digital Education Action Plan (European Commission & OECD, 2025), and the next generation of cognitive offloading studies, becomes available.

Four specific empirical priorities follow. First, longitudinal cohort studies on developmental cognitive consequences of GenAI exposure across the K-12 window, with appropriate controls for socioeconomic status, pre-existing cognitive variation, and intensity of exposure. Second, controlled evaluation of the 4D framework's specific developmental sequencing claims, testing whether 4D-trained cohorts demonstrate superior Discernment, Diligence, and operational fluency outcomes. Third, cross-cultural validation of the framework's applicability across diverse educational contexts and underrepresented linguistic communities. Fourth, development of institutional vendor-evaluation frameworks that operationalise methodology transparency, regulatory cooperation, and safety research investment as comparable, auditable institutional decision inputs.

Three further methodological priorities are worth foregrounding. First, the empirical literature reviewed in Section 3 relies heavily on short-duration experimental designs, often single-session essay tasks under controlled conditions. The cognitive routines the 4D framework targets are formed across years of practice, not minutes of laboratory engagement. The research designs that can detect framework-relevant effects at developmental timescales are longitudinal, expensive, and methodologically demanding. They are also essential. Without them, the field will continue to mistake short-term task-performance effects for long-term routine-formation effects, with serious consequences for institutional decision-making. Second, the field has under-invested in qualitative and ethnographic work documenting what AI use actually looks like in classrooms, both as instructed and as practised. The ethnographic record will become important historical material as the technology generation shifts; we should be building it now. Third, the empirical literature is currently weighted toward studies of intervention rather than studies of non-intervention. We know more about what happens when students are taught to use AI in a particular way than about what happens when they are not. The control condition (unguided default use) is the dominant student experience and is methodologically under-examined.

Finally, a note on the present author's positionality. The framework offered here is developed by a single author with operational teaching experience across Anglo-American, Czech, and US-affiliated tertiary institutions and consulting experience

across multiple AI-deploying industries. It is not the product of a research group, a multi-institution collaboration, or a funded research programme. Its strengths are coherence, operational specificity, and currency. Its weaknesses include the limitations of a single perspective, the absence of formal peer review at this stage, and the specific cultural and institutional biases of the author's professional trajectory. Readers should weight the recommendations accordingly. The framework is offered as a serious contribution to the field's working conversation, not as a finished research product. Constructive engagement, methodological critique, and empirical examination of the framework's claims are explicitly invited.

13. Conclusion

The transition currently underway in education is not principally about a technology. It is about whether educational institutions can evolve fast enough, intelligently enough, and humanely enough to remain the institutions through which the next generation forms the cognitive capacities a free society requires. Frontier generative AI is, on one reading, the most powerful cognitive partner the species has ever invented. On another, it is the most powerful cognitive substitute. The pedagogical question is which it becomes for the child now in primary school, the adolescent now in secondary, the young adult now in tertiary education. The answer depends on what we teach them about it, in what order, with what protection, in service of what cognitive capacity.

The 4D framework (Delegation, Description, Discernment, Diligence) offers one structured way to teach AI fluency as a graduated competence rather than an undifferentiated permission. The developmental tiering proposed here (protective AI-adjacent literacy in primary education, scaffolded AI literacy with explicit friction in secondary, operational AI fluency in tertiary) offers one structured way to align that competence with the developing brain. The model-selection criteria offered (methodology transparency, regulatory cooperation, safety research investment) offer one structured way for institutions to exercise fiduciary judgement. The international policy frameworks now converging (EU AI Act, UNESCO competency frameworks, EU-OECD AILit, PISA 2029 MAIL) provide the competency outcome architecture against which institutional reform can be calibrated.

None of these proposals is final. All are working hypotheses, intended to be revised as evidence accumulates and as the technology continues to evolve. What I am confident about is that the current dominant postures (prohibition, unregulated adoption, surveillance-based compliance) are each, in their own way, developmental illiteracy. They cannot survive the decade ahead. Institutions that begin operational reform now, with deliberate curriculum redesign, educator preparation, assessment redesign, and protection-by-design infrastructure, will be positioned to serve their learners. Institutions that defer reform will pay rising costs to catch up under rising external pressure. Teaching the next generation to think with AI rather than around it requires educators willing to do the same.

References

- AI Literacy Institute. (2026, March 3). AI Literacy Review – March 3, 2026. <https://ailiteracy.institute/ai-literacy-review-march-3-2026/>
- Anthropic. (2025). AI Fluency for Educators [Course]. Anthropic. <https://www.anthropic.com/learn>
- Anthropic. (2026). Responsible Scaling Policy, version 3.0. Anthropic. <https://www.anthropic.com/responsible-scaling-policy>
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073. <https://doi.org/10.48550/arXiv.2212.08073>
- California Department of Education. (2025). Artificial Intelligence in Education: Guidance for K-12 Schools. California Department of Education. <https://www.cde.ca.gov/ci/pl/aiincalifornia.asp>
- Chew, H. E., Soon, C., & Kaur, H. (2026). Unified AI Literacy Framework for Singapore. Institute of Policy Studies, National University of Singapore.
- Cunningham, C. J. L., et al. (2025). One year in the classroom with ChatGPT: empirical insights and transformative impacts. *Frontiers in Education*, 10, 1574477. <https://doi.org/10.3389/feduc.2025.1574477>
- Deng, R., Wang, X., & Zhang, Y. (2026). ChatGPT's impact on student learning outcomes: A meta-analysis of 35 experimental studies. *Humanities and Social Sciences Communications*, advance online publication. <https://doi.org/10.1057/s41599-026-07019-z>
- Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, 64, 135–168. <https://doi.org/10.1146/annurev-psych-113011-143750>
- Digital Promise. (2024). Revealing an AI Literacy Framework for Learners and Educators. Digital Promise. <https://digitalpromise.org/2024/02/21/revealing-an-ai-literacy-framework-for-learners-and-educators/>
- Ergene, Ö., & Ergene, B. Ç. (2025). Cognitive offloading and over-dependence on generative AI in higher education. *Education and Information Technologies*, advance online publication.
- European Commission, & OECD. (2025). Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education [Review draft]. European Commission & OECD. <https://ailiteracyframework.org/>
- European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 1689/1.
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489–530.
- Gascón-Cánovas, J. J., et al. (2025). Executive function in adolescents with problematic smartphone use. *International Journal of Adolescence and Youth*, 30(1), Article 2562950. <https://doi.org/10.1080/02673843.2025.2562950>
- 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>
- Gilbert, S. J. (2024). Cognitive offloading is value-based decision making: Modelling cognitive effort and the expected value of memory. *Cognition*, 247, 105783.
- Hinze, V., Blakemore, S.-J., Dalgleish, T., Ford, T., Mansfield, K. L., Ukoumunne, O. C., Kuyken, W., & Montero-Marin, J. (2025). The longitudinal relationship between self-reported executive function and mental health in early adolescence. *Journal of the American Academy of Child & Adolescent Psychiatry – Open*, advance online publication. <https://doi.org/10.1016/j.jaacop.2025.07.003>
- Hutton, J. S., Dudley, J., Horowitz-Kraus, T., et al. (2022). Associations between screen-based media use and brain white matter integrity in preschool-aged children. *JAMA Pediatrics*, advance online publication.

- Illingworth, S. (2026). What is critical AI literacy? AI Literacy Institute Working Papers.
- Kharbach, M. (2026). UNESCO's AI Competency Framework for Students: What every educator should know. medkharbach.com.
- Kooli, C., et al. (2025). Generative AI Addiction Syndrome: A conceptual framework. *Computers in Human Behavior Reports*, advance online publication.
- Kosmyna, N., Hauptmann, S. C., Yuan, T., Situ, Y., 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 tasks. MIT Media Lab Technical Report. <https://doi.org/10.48550/arXiv.2506.08872>
- Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2024). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence*, 7, 100101.
- Lee, J., & Low, V. (2024). Generative AI in academic writing: Trade-offs between efficiency and originality. *Computers & Education: Artificial Intelligence*, 7, 100268.
- León-Domínguez, U. (2024). Potential cognitive risks of generative transformer-based AI chatbots on higher-order executive functions. *Neuropsychology*, 38(4), 293–308. <https://doi.org/10.1037/neu0000948>
- Lewin, K. M., et al. (2025). The longitudinal impact of screen media activities on brain function, architecture and mental health in early adolescence. *International Journal of Clinical and Health Psychology*, 25(2).
- Lin, Z., Dai, Y., & Ng, O.-L. (2025). Constructionism in K-12 AI literacy education: A systematic review of pedagogical designs, student outcomes, and learning mechanisms. *Journal of Educational Computing Research*, advance online publication. <https://doi.org/10.1177/07356331251360442>
- Martin, F., Bacak, J., Polly, D., & Wang, W. (2024). A systematic review of research on K-12 AI literacy. *Computers and Education: Artificial Intelligence*, 7, 100295.
- Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO.
- OECD. (2026). PISA 2029 Media and Artificial Intelligence Literacy (MAIL) Assessment Framework: Navigating an Evolving Digital World [Draft]. OECD Publishing. <https://www.oecd.org/en/about/projects/pisa-2029-media-and-artificial-intelligence-literacy.html>
- Perrigo, B. (2026, February 24). Exclusive: Anthropic drops flagship safety pledge. TIME. <https://time.com/7380854/exclusive-anthropic-drops-flagship-safety-pledge/>
- Raptopoulou, A. (2025). ChatGPT in higher education: Supporting academic literacy through ChatGPT-based activities. *European Journal of Education*, e70131. <https://doi.org/10.1111/ejed.70131>
- Reggente, N., et al. (2026). The impact of smartphone use on brain function in adolescence: A scoping review. *Pediatric Reports*, 18(2), 43. <https://doi.org/10.3390/pediatric18020043>
- Roe, J., Furze, L., & Perkins, M. (2026). Digital plastic: A metaphorical framework for critical AI literacy in the multiliteracies era. *Postdigital Science and Education*, advance online publication.
- Steinberg, L. (2008). A social neuroscience perspective on adolescent risk-taking. *Developmental Review*, 28(1), 78–106. <https://doi.org/10.1016/j.dr.2007.08.002>
- Su, J., Ng, D. T. K., & Chu, S. K. W. (2025). Defining, enhancing, and assessing artificial intelligence literacy and competency in K-12 education from a systematic review. *Interactive Learning Environments*, 33(9), 5766–5788. <https://doi.org/10.1080/10494820.2025.2487538>
- Thomas, J. W. (2000). A review of research on project-based learning. Autodesk Foundation.
- UNESCO. (2024). AI competency framework for students and teachers. UNESCO Publishing.
- Welker, M. (2025). Global AI literacy readiness assessment 2025. UNESCO/UNICEF Working Paper.
- Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33–35. <https://doi.org/10.1145/1118178.1118215>

- Xing, B., et al. (2025). ChatGPT in K-12 education: Effects on student learning outcomes. *Journal of Educational Computing Research*, advance online publication.
- Zhang, S., Xiao, R., Botelho, A. F., Liao, G., Chiu, T. K. F., Stamper, J., & Koedinger, K. R. (2026). How to assess AI literacy: Misalignment between self-reported and objective-based measures. *arXiv preprint arXiv:2601.06101*.

© 2026 N.J. Berky. This working paper is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0). You are free to share and adapt with appropriate attribution. Permanent identifier · ORCID 0009-0001-2227-215X. Suggested citation · Berky, N.J. (2026). *Teaching With, Not Around: A Developmental Framework for Generative AI Integration in Primary, Secondary, and Tertiary Education (2023–2026)*. Working paper, version 2.0. NJB Berky Advisory.