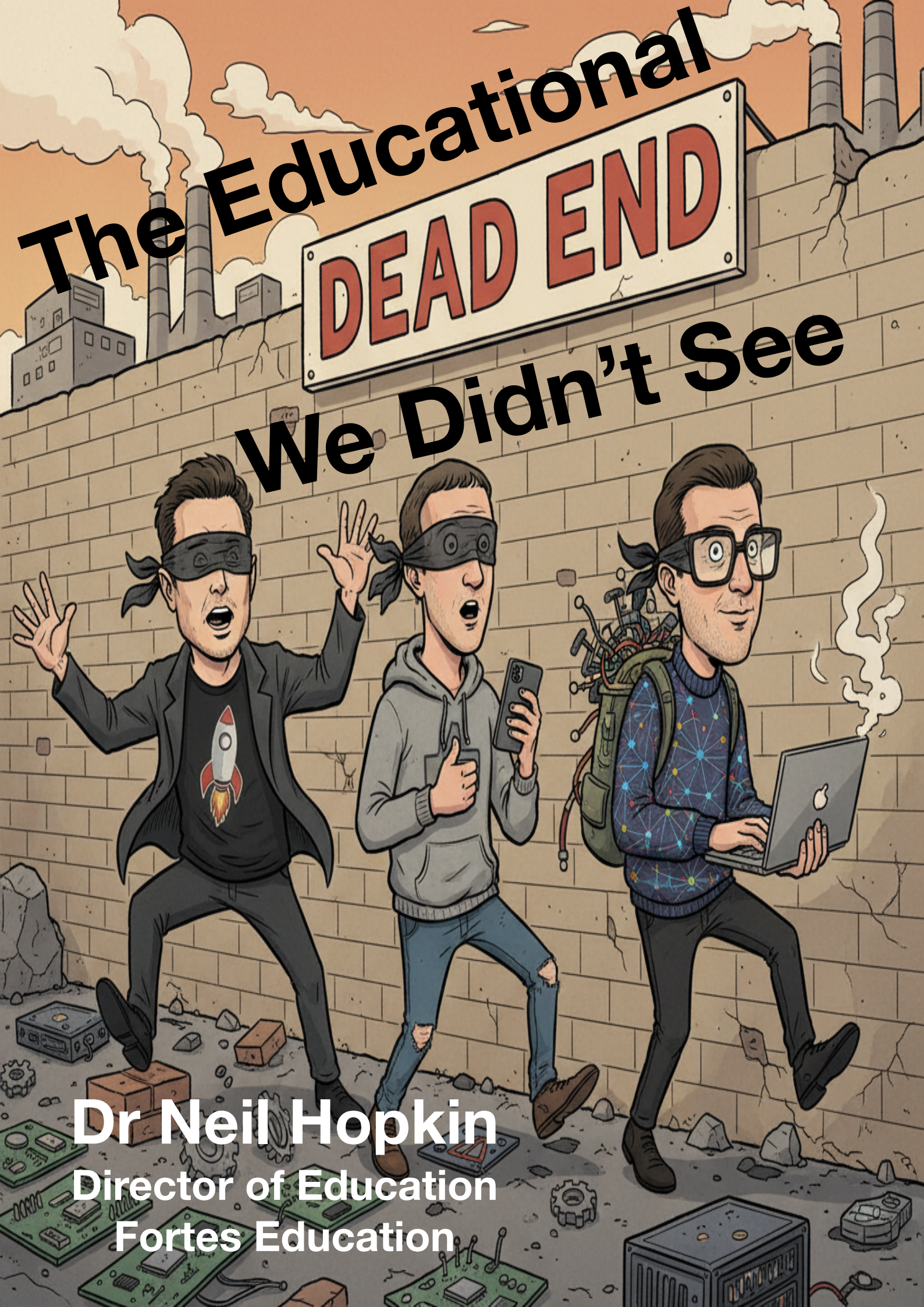
The Educational Dead End We Didn't See



Dr Neil Hopkin
Director of Education
Fortes Education

Introduction

On a Thursday morning in early spring, in a London comprehensive, a science teacher stood over a neat stack of Year 9 essays. For a brief moment, she felt the flicker of relief teachers crave: no spilled ink, no half-finished scribbles, no tortured grammar. Each essay was tidy, perfectly formatted, and suspiciously well-organised. Topic sentences lined up like soldiers. Arguments marched to conclusion. She lingered on one student whose work had never quite found its footing in previous terms. Now, suddenly, the prose flowed. It was the kind of improvement schools used to throw celebrations for.

And yet, her relief curdled almost immediately. Something was missing—the wrestling, the sense of thought in motion. The voice on the page sounded like her student, but stripped of hesitation, stripped of originality, stripped of the little stumbles that signal the brain stretching into new territory. It was too good. Too even. Too borrowed.

This sense of hollowness has become familiar across classrooms from Dubai to Dublin. Teachers aren't celebrating polished AI-assisted work; they're unsettled by it. They sense that the cognitive labour of learning, holding an idea in mind, comparing one claim to another, turning a concept around to test its edges, has been displaced by the simulation of that labour. What remains is the residue of thinking without the thinking itself.

Psychologists have a name for this tendency: cognitive offloading. When people expect information to be digitally stored and accessible, they retain less of the content itself and more of where to find it. Sparrow, Liu and Wegner (2011) demonstrated the effect in a series of experiments, showing how internet access subtly altered not just what participants remembered but how they remembered it. It was not knowledge itself that stuck, but the pathway to retrieval. Risko and Gilbert (2016) extended this argument, showing that offloading was not limited to memory but extended to attention, planning, even problem-solving. When tools are available, people unconsciously reorganise their tasks so that the tool carries more of the strain.

This is not new. Wegner (1987) coined the concept of transactive memory to describe how groups distribute knowledge: one person remembers names, another dates, another stories. Together, they form a composite memory stronger than any individual's. The difference is that today the group member is not a person but a predictive machine. The partner in our transactive memory does not merely recall; it anticipates, completes, and presents.

In classrooms, this dynamic is not abstract. Teachers see it when a student breezes through a written assignment with unexpected polish, only to flounder in oral questioning. They see it when homework is meticulous, but in-class reasoning is fragile. They see it when a group discussion reveals that what seemed like a leap in learning was only a leap in presentation. Students are outsourcing the messy, formative struggle that constitutes real understanding.

But here's where the puzzle thickens. The problem isn't simply that students are using machines as shortcuts. The deeper problem is what happens when you push past surface polish. Take the ecosystem essay again. The student writes fluently about photosynthesis, energy transfer, and trophic levels. But when the teacher asks: What happens if rainfall drops by thirty percent for two consecutive years?: everything buckles. The prose continues, but the reasoning collapses. The student, leaning on AI-generated explanations, produces elegant sentences that misrepresent cascades, confuse cause and effect, and stumble over the basic logic of resource scarcity.

This brittleness repeats across subjects. In maths, large language models are dazzling at reciting algebraic procedures but falter when a word problem shifts context slightly (Madaan et al., 2023). In history, they can deliver competent summaries of the causes of the Cold War but flail when asked to reason through a counterfactual, say, a 1990s Europe where the Berlin Wall never fell. In literature, they can identify themes in Hamlet but misuse them in comparative essays where interpretation must adapt.

These failures are not incidental. They go to the heart of what education is supposed to achieve: transfer. The ability to take a concept learned in one situation and apply it in another is what distinguishes learning from parroting. It is also, ironically, the very thing AI most often fails to do.

Now comes the mystery. If these systems can ace exams, pass professional tests, and produce outputs that dazzle in familiar conditions, why do they collapse in the ordinary novelty of the classroom? Why can a machine that composes a polished essay on ecosystems not follow the causal chain of a drought? Why can a student armed with this machine not transfer knowledge across a trivial twist?

The answer cannot simply be “students are cheating” or “teachers aren’t adapting fast enough.” The brittleness runs deeper. It is not that we are using the tools wrongly, but that the tools themselves may be optimised for the wrong thing. And that possibility is hard to face. If true, it means that the map Silicon Valley has given us, the one promising that more data and bigger models will carry us steadily toward intelligence, may not be a map of the territory schools actually need.

Surely not. Surely the paradigm cannot be that wrong.

The Stubbed Toe of Our Current Direction with AI in Education

The recognition comes, often, like a stubbed toe. You walk down a corridor you’ve walked a hundred times, thinking about something else, and suddenly pain brings you to attention. Something was there you hadn’t noticed. The obstacle was always there, but invisible until contact. For education, the stubbed toe is the moment the surface promise of AI collides with its structural limits. A maths teacher discovers that a ratio problem, slightly rephrased, yields confident nonsense. A science teacher watches an AI lab report collapse under a changed variable. A history teacher finds that a counterfactual question elicits platitudes where reasoning should be. Each moment feels like an anomaly, until you see the pattern.

Researchers, too, have begun cataloguing the bruises. In 2024, MIT’s Computer Science and Artificial Intelligence Laboratory published a study showing that the reasoning skills of large language models were regularly overestimated. When evaluated on counterfactual tasks, (problems that ask “what if something had been different?”), the models faltered (MIT CSAIL, 2024). This isn’t a niche test; counterfactual reasoning is a bedrock of education. Every “what if” in history, every “if X then Y” in science, every hypothetical in literature relies on it.

Parallel research has pressed on another fault line: compositional generalisation, the ability to recombine known parts into new wholes. Yang (2024) tested language models on tasks that required assembling familiar components in unfamiliar configurations. Even strong models stumbled. The same bricks that built coherent sentences failed to build coherent reasoning when rearranged. And then came a more stinging revelation: some of the “emergent abilities” of large models may not be abilities at all but artefacts of how we measure them. Schaeffer, Miranda and Koyejo (2023) argued that discontinuities in performance charts may simply be illusions produced by the quirks of evaluation metrics. What looked like sudden leaps in reasoning could be statistical mirages. Put together, these findings form a pattern teachers already suspect. The models are not stumbling on obscure brainteasers; they are faltering on exactly the kinds of novelty education requires. Fluency in distribution does not guarantee competence out of distribution.

Teachers live this every day. A Year 8 maths teacher in Manchester asked her students, with AI assistance, to explain ratios using three different contexts. The answers came back grammatically correct and well-structured, but two contained subtle errors, misapplied formulas, misplaced terms, that betrayed a lack of underlying grasp. “It’s like being gaslit by a very polite person,” she joked in the staffroom. The joke masked an unease: if the AI could produce convincing nonsense, how were students to learn the difference? An English professor at a U.S. university reported something similar. AI-generated essays scored highly on rubrics that valued structure and clarity but consistently underperformed in originality and depth (Schaeffer et al., 2024). The essays were, in a sense, optimised for grading criteria rather than for thought.

These are the bruises on the collective toes of educators. They reveal that what looks like progress may in fact be mismeasurement. The “state-of-the-art” benchmarks, the leaderboards,

the glowing slides in procurement meetings: they capture fluency in known grooves, not the ability to travel across new terrain. At first, each stubbed toe feels like an inconvenience. But accumulate enough of them, and a dawning horror begins to set in: what if this isn't incidental? What if collapse under novelty is a feature of the paradigm, not a bug? If that's true, then our problem in education is not just how students use AI but how the very design philosophy of AI mismatches the goals of schooling. We are optimising machines for surface competence and expecting them to deliver transferable understanding. It is as though we are marching confidently down a path that leads not to deeper learning but to a cul-de-sac.

The Dead Language of Progress

Every era of education falls under the spell of a new vocabulary. In the 1960s, it was "programmed learning": Skinner boxes, linear teaching machines, and the promise that drills plus reinforcement could replace the craft of teaching. By the 1980s, it was "computer literacy": lessons built around floppy disks, early word processors, and the assumption that exposure alone would future-proof students. In the 2000s, it was "digital natives" and "21st-century skills": slogans that crowded conference stages but often left classrooms puzzled about what was meant.

Today, the language comes not from education ministries but from Silicon Valley. Tokens. Benchmarks. Parameters. Context windows. State-of-the-art performance. The words circulate with the confidence of neutrality. A procurement officer in Dubai hears them in a vendor pitch; a superintendent in Toronto repeats them in a strategy memo; a headteacher in Singapore hears them repackaged in a professional development workshop. The language sounds empirical, but it carries an ideology. The ideology is this: intelligence is a slope you climb by feeding a system more data. Understanding is measured by the number of tasks on which fluency improves. If accuracy climbs on benchmarks, then progress must be happening. If context windows expand, then reasoning must be broadening.

But history teaches caution. Education has always been haunted by the danger of measuring the wrong thing with great precision. In the 19th century, inspectors judged handwriting slant and uniformity as proxies for comprehension. In the 20th, speed of recall was mistaken for depth of understanding. We know how easily the sheen of performance can be confused with the substance of learning.

The AI world now risks repeating the same mistake. Schaeffer, Miranda and Koyejo (2023) showed that some of the most celebrated "emergent abilities" of large language models may be artefacts of measurement. Sudden jumps in charts, interpreted as breakthroughs in reasoning, were sometimes illusions produced by discontinuous evaluation metrics. Smooth, shallow curves were sliced into jagged leaps, creating the illusion of intelligence where there was only statistical artefact.

Meanwhile, cognitive scientists have been telling a different story. Human learning is not a product of massive exposure. It is sparse, structured, and compositional. Tenenbaum and colleagues (2011) argued that children generalise from very few examples because they represent the world in causal frameworks. Lake et al. (2017) showed that humans can recombine familiar concepts in novel ways: a process machines still struggle with. You do not need a million examples of a cup falling from a table to infer gravity. You need to understand causality itself. Yejin Choi calls this commonsense the "dark matter of intelligence" (Choi, 2023). Like its cosmological counterpart, it is invisible to measurement but shapes everything that matters. A model without commonsense can produce fluent prose but collapse under the smallest change of conditions.

This is why teachers feel a growing mismatch. A Year 6 pupil who has learned ratio should be able to transfer it to cooking recipes, scaling maps, and mixing chemicals. A Year 9 student who understands ecosystems should predict what happens when rainfall drops. A Year 11 who learns to spot logical fallacies should recognise them in political speeches as easily as in classroom debates. Schools exist to cultivate these transfers. Yet the language borrowed from Silicon Valley has little space for transfer. It measures surface fluency in familiar grooves, not the causal structures that travel.

This is how a dead language hardens into common sense. We reward polish, mistake benchmarks for depth, and let procurement rubrics shape pedagogy. The slide decks glow green while the classrooms grow hollow.

The Dawning Horror

The hinge of this story arrives from an unexpected quarter: the career of Song-Chun Zhu. For thirty years, Zhu built his reputation across Brown, Harvard, and UCLA. Unlike many in the deep learning mainstream, his focus was not on scaling data but on uncovering the structures of intelligence. With David Mumford, he developed a stochastic grammar of images, a way of representing visual scenes not as pixels but as hierarchical parts and relations (Zhu & Mumford, 2006). Later, he explored Bayesian models of vision that combined perception with causal reasoning (Zhu et al., 2009). His research pointed again and again to the same claim: intelligence depends on representations that capture structure, not just frequency.

By 2020, Zhu had concluded that the prevailing paradigm, scale up the models, climb the benchmarks, would not deliver general intelligence. He founded the Beijing Institute for General Artificial Intelligence (BIGAI) to pursue an alternative. There, his team proposed the Tong Test (Peng et al., 2023). Unlike the Turing Test, which asks whether a machine can imitate human conversation, the Tong Test evaluates whether an agent can succeed in dynamic embodied physical and social interactions (DEPSI). Can it pursue a goal in a changing environment? Can it revise a plan when conditions shift? Can it learn from sparse data and articulate the reasoning behind its choices?

For educators, this sounds familiar. A Year 4 child builds a circuit. The bulb doesn't light. She revises the connections and explains why. A Year 10 student writes a thesis, encounters counterargument, and refines her essay. A Year 9 group models an ecosystem, realises predictions fail under new rainfall conditions, and revises the causal chain. These are not acts of brute force; they are acts of adaptation under constraint. Zhu's work crystallises the dawning horror: the failures teachers observe are not accidental side effects. They are baked into the paradigm. Systems optimised for big data, small task will always excel at surface fluency and fail at transfer. Schools exist to do the opposite.

This mismatch explains why a student's AI-assisted essay collapses under questioning, why polished outputs hide brittle reasoning, why counterfactuals expose hollow understanding. It explains why benchmarks glow while classrooms stumble. The problem is not students, nor even teachers, but the map itself. We stand now at the edge of the cul-de-sac. Behind us: years of investment in tools tuned to benchmarks, outputs, and fluency. Ahead: an uncharted road mapped by those who insist intelligence is sparse, causal, embodied.

Imagine again the London science classroom. The teacher asks: What happens if rainfall drops by thirty percent for two consecutive years? Under the old paradigm, the machine produces prose: fluent, confident, but brittle. Under the new, it sketches a causal model: reduced rainfall weakens primary producers, destabilises higher trophic levels, triggers trophic cascades. It tests the prediction against ecological principles, explains the reasoning, and adapts when pushed. The student doesn't copy; she interrogates. The rubric has shifted too: alongside "clarity" and "accuracy," there are rows for Explainability and Counterfactual Robustness.

This is what "small data, big task" looks like when translated into pedagogy. Procurement officers ask not how many parameters a model has, but whether it can learn from a handful of classroom examples. Policymakers stop measuring success in terms of productivity and start valuing adaptability under constraint. Teachers stop fearing AI as cognitive offloading and start using it as a scaffold for reasoning. But before we reach that road, we must acknowledge where we are. The language of tokens and scale has led us into a dead end. We mistook polish for thought, speed for understanding, benchmarks for transfer. If we continue, classrooms will keep stubbing toes on invisible steps. The problem is not just how we use the machines. The problem is the paradigm itself.

The Sinking Ship: Song-Chun Zhu's Departure

For most of his career, Song-Chun Zhu worked inside the very institutions that defined American artificial intelligence. His CV reads like a map of the U.S. academy: Brown for graduate study, Harvard for postdoctoral research, then nearly three decades at UCLA, where his lab became a crucible for students interested in computer vision, probability theory, and cognition. By the time he left, Zhu was not a marginal figure; he was a central one. He held endowed chairs, led major research centres, and published in the most prestigious venues of computer science.

And yet, in 2020, he stepped away. To colleagues, the move looked curious, even inexplicable. Why would a tenured professor, embedded in one of the most powerful scientific systems in the world, choose to leave? Why exchange a well-funded lab in California for the uncertainties of building something new in Beijing? The answer is not found in geopolitics, but in ideas. Zhu was convinced that the path Silicon Valley had chosen, bigger datasets, larger models, faster hardware, was not the path to general intelligence. He had been saying so for years, often politely, sometimes bluntly. Scaling, he argued, produced fluency but not understanding. His departure from UCLA was not a rejection of a place but of a paradigm.

A Different Lineage

To understand Zhu's decision, it helps to see the lineage of his work. His doctoral studies brought him into contact with David Mumford, a Fields Medal–winning mathematician who had turned his attention to vision and pattern recognition. Together, they developed the idea of a stochastic grammar of images (Zhu & Mumford, 2006). The insight was deceptively simple: just as language has grammar, rules for combining words into sentences, images have grammar too. An image is not a cloud of pixels; it is a structured arrangement of parts and relations.

In their framework, an image of a chair, for example, could be parsed into legs, seat, and backrest, each with probabilistic rules about how they combine. A scene could be represented as a composition of such objects, themselves nested in higher-order structures. The result was a model that did not just classify images but explained them. It could say not only "this is a chair" but also "these are the parts and relations that make it a chair."

Later, Zhu and colleagues extended this into AND-OR graphs (Zhu, Luo & Wu, 2009). These models represented knowledge as a tree of possibilities: AND nodes for structures that must co-occur, OR nodes for alternatives. An object could be represented as a set of mandatory components (AND) with optional variations (OR). This framework allowed machines to generate new instances, parse ambiguous data, and reason about causes. These were not just technical contributions. They reflected a worldview: intelligence is not the accumulation of patterns but the ability to represent and manipulate structure. Where the mainstream of deep learning was drifting toward ever-larger convolutional nets trained on millions of images, Zhu's models tried to capture the grammar behind appearances.

By the late 2010s, the dissonance between Zhu's worldview and the industry mainstream had become acute. Tech companies were scaling models to billions of parameters. OpenAI's GPT-2 (Radford et al., 2019) stunned the field with its ability to generate humanlike text; GPT-3 (Brown et al., 2020) multiplied that ability by orders of magnitude. Google's PaLM (Chowdhery et al., 2022) and Anthropic's Claude followed, each leap powered not by new theories but by more data, more compute, more scale.

The success was undeniable. Benchmarks fell like dominoes: summarisation, translation, question answering, even code generation. To most, this was proof that scale was the royal road to intelligence. But to Zhu, it was proof of something else: the field had mistaken fluency for understanding. He pointed to brittleness. Models that aced exams often collapsed when tasks were reframed even slightly (Talmor et al., 2023). Systems that generated eloquent essays faltered when asked to reason counterfactually (Kosinski, 2023). Models that mastered algebraic procedures stumbled on novel word problems (Madaan et al., 2023). The failures were not random; they clustered precisely where structure and causality mattered. In lectures, Zhu would sketch the contrast: humans generalise from tiny data by inferring rules. A child who sees a few

examples of dogs can recognise a new one, not because she has memorised features but because she grasps abstract properties—legs, tail, bark. Machines trained on millions of images that still fail to make such a leap, Zhu argued, are on the wrong path.

Leaving the Hall

There is an image that fits the moment. Imagine a conference hall where everyone is watching the same chart climb: benchmark scores rising with scale. The applause is genuine, the excitement real. But one figure leaves the hall, stepping into a quiet courtyard. To those inside, it looks like retreat. To the one who left, it is a search for clearer air.

That courtyard was literal for Zhu. At Peking University, near Weiming Lake, he established the Beijing Institute for General Artificial Intelligence (BIGAI). Students describe him pacing the courtyard with chalk in hand, sketching diagrams of causal models on blackboards, asking questions not about how to scale but about how to represent. He described his mission in simple terms: “Over the last 30 years, I’ve been focused on one thing: to build understanding” (Peng et al., 2023). The words were almost embarrassingly plain. But they cut against the grain of the field. While others spoke of petaflops, training runs, and parameter counts, Zhu spoke of understanding.

To colleagues in the U.S., Zhu’s move was puzzling, even risky. To him, it was necessary. The institutions most flush with capital, OpenAI, Google, Meta, had locked into the scaling paradigm. Academic labs, increasingly dependent on corporate funding, followed suit. To pursue an alternative vision, Zhu needed a different environment. In this sense, his departure was less an exile than a rebirth. BIGAI became a space where causal reasoning, compositionality, and small-data learning could be foregrounded. His team proposed a new evaluation, the Tong Test (Peng et al., 2023), designed to measure intelligence not by static benchmarks but by performance in dynamic, embodied, interactive settings.

The symbolism of the courtyard deepened. Exiled from the noise of benchmark races, Zhu cultivated a new language of intelligence: sparse data, causal models, transfer under constraint. What looked eccentric was, in fact, a hinge.

For educators, the resonance is striking. Schools, too, have been seduced by fluency metrics. In many systems, assessment has been narrowed to standardised tests, multiple-choice items, and rubrics that reward polish. Ministries celebrate rising scores as proof of learning. But teachers see the brittleness: students who ace exams but cannot transfer knowledge to new contexts; essays that look polished but collapse under questioning. The story of Zhu’s departure offers a mirror. Just as Silicon Valley mistook fluency for intelligence, education risks mistaking performance metrics for learning. Just as Zhu left the hall of benchmarks for the courtyard of understanding, schools may need to step away from surface measures toward deeper ones.

This pattern, mistaking fluency for depth, surface for substance, is not new. In the mid-20th century, B. F. Skinner promoted “teaching machines” that delivered programmed instruction through mechanical devices. Students progressed through linear frames, receiving reinforcement for correct answers. The machines produced measurable gains in speed and accuracy. But critics noted that they trained rote responses, not transfer or understanding (Suppes, 1966). By the 1970s, the enthusiasm had waned, leaving behind the cautionary tale of mistaking measurable performance for learning.

Zhu’s critique echoes another intellectual battle: Noam Chomsky’s challenge to behaviourism. Chomsky (1959) argued that children’s ability to produce novel sentences could not be explained by reinforcement of observed patterns. The “poverty of the stimulus” meant that learning required internal structure, not just exposure. Chomsky’s critique redirected linguistics toward generative grammar and cognitive science. Zhu’s move, likewise, signals a pivot: from data accumulation to structure building. His departure forces a question that education, too, must ask: are we following the right map, or are we mistaking fluency for thought?

Big Data, Small Task vs. Small Data, Big Task

The story of scale has always been beguiling. It speaks to something deep in the modern imagination: the belief that more is better, that size itself carries a kind of inevitability. When the first massive language models emerged, GPT-2, GPT-3, the numbers alone were intoxicating. Billions of parameters, trained on oceans of text scraped from the internet, producing prose that sounded like it had been written by people. To many, it felt like we were glimpsing the early steps of intelligence itself.

The story was reinforced by the graphs. Accuracy curves that had plateaued for years suddenly bent upward. Benchmarks that once seemed resistant, machine translation, summarisation, even logical reasoning puzzles, were overtaken. Each release was accompanied by glossy reports: larger models performed better across almost every metric. The message was simple: keep scaling, and general intelligence will come.

But here is the twist. The very success of these systems concealed a fragility that was visible only when you changed the angle of the light. In 2024, researchers at MIT's CSAIL tilted familiar tasks into counterfactual form and watched performance crumble (MIT CSAIL, 2024). A model that could answer a science question about ecosystems failed when rainfall variables were altered. A system that produced convincing summaries of historical events faltered when asked what might have happened if one variable changed. The surface fluency remained; the reasoning collapsed.

The brittleness is not limited to counterfactuals. Work on compositional generalisation has shown that even when models master every brick in the wall, they struggle to build with those bricks in new configurations (Yang, 2024). A system that knows every individual concept may still fail to combine them coherently. In classrooms, this plays out in ways teachers recognise immediately. A Year 8 student can use an AI tutor to solve rehearsed algebraic equations, but when the same variables are embedded in a word problem about dividing recipes or calculating speed, the student, and the model, flounder.

This is the paradox of big data, small task. Models trained on vast datasets become highly competent within distribution but fragile when asked to step outside. They are optimised for narrow grooves of performance, but real intelligence, human intelligence, reveals itself in transfer, in the ability to adapt when the groove shifts.

Zhu's alternative, what he called small data, big task, inverts the frame. Instead of asking how far a system can go with oceans of data, the challenge is to ask how much it can do with very little. Can it learn from five examples? Can it build a causal model that travels across situations? Can it adapt when conditions change and explain why? This may sound counterintuitive, but it mirrors the way humans actually learn. Children do not need to see a million dogs to understand what makes a dog. They need only a few encounters to grasp the underlying concept: four legs, tail, bark. Once the structure is grasped, the category is robust. A toddler who has never seen a Dalmatian can still recognise it as a dog.

Psychologists and cognitive scientists have been describing this phenomenon for decades. Tenenbaum, Kemp, Griffiths and Goodman (2011) argued that human learning relies on Bayesian inference—structured hypotheses that generalise far beyond the data observed. Lake, Ullman, Tenenbaum and Gershman (2017) showed how humans construct new concepts by recombining familiar parts, generating original ideas from minimal input. Yejin Choi (2023) described commonsense as the "dark matter of intelligence": invisible to measurement, but indispensable to reasoning. And Gary Marcus has argued repeatedly that without explicit structure, rules, causal models, representations, scale produces only "fluent errors" (Marcus & Davis, 2019). The evidence converges on a single point: more data can improve fluency, but fluency is not the same as understanding. Understanding requires models that capture structure. Without them, systems produce eloquence without depth, polish without transfer.

Consider the classroom analogies. A Year 11 student submits a flawless essay on Macbeth, paragraphs arranged with textbook precision. Yet in a seminar, when asked to compare Macbeth's ambition with another character from a different play, she falters. The AI-assisted polish did not translate into reasoning. Or take mathematics. A student completes a worksheet of

quadratic equations flawlessly, but when the problem is reframed into a real-world scenario, say, calculating the trajectory of a ball, the solution unravels. The skills acquired were narrow; the transfer was missing. This is not a failing of students so much as a reflection of the paradigm they are being trained within. Big data, small task produces competence in grooves. Education, by contrast, is supposed to cultivate adaptability. When schools adopt AI tools optimised for fluency, they risk aligning themselves with the very brittleness they are meant to overcome.

The danger is compounded by policy. Once benchmarks dominate, the system orients itself around them. In AI, leaderboards drive research agendas. In schools, standardised tests drive curriculum. Both produce the same illusion: rising scores as proof of progress, even as underlying robustness stagnates. Policymakers celebrate outputs, inspection frameworks reward polish, and procurement teams look for glowing charts. The cycle feeds itself.

Zhu's small-data, big-task paradigm suggests another way. Instead of measuring whether a system can produce the right answer under familiar conditions, measure whether it can adapt under new ones. Instead of rewarding polish, reward transfer. Instead of training machines, and students, to replicate, train them to model, explain, and revise. The implications are profound. Imagine a history exam that includes not only "What were the causes of the Cold War?" but also "What if the Berlin Wall had not fallen in 1989?" The first question rewards recall and fluency; the second tests transfer and causal reasoning. Imagine a science assessment where a student builds a model of an ecosystem, then has to adjust it under changing conditions: drought, invasive species, new policies. Imagine a maths exam where the equations are familiar but the contexts shift, requiring reasoning rather than rote. The pattern here is the same as in Zhu's research. Intelligence is revealed not in smooth reproduction but in resilience under constraint. The small-data, big-task frame forces both machines and students to show their models, not just their answers.

History offers echoes of this struggle. In the 1960s, B.F. Skinner's "teaching machines" promised efficiency by delivering programmed instruction in linear frames. The outputs improved: students answered faster and more accurately. But the learning was brittle. Transfer was absent. By the 1970s, enthusiasm waned, leaving a cautionary tale about mistaking speed for depth (Suppes, 1966). A decade later, Benjamin Bloom's "2 sigma problem" demonstrated that one-to-one tutoring could produce dramatic gains in performance, but the insight was often reduced to surface metrics—test scores—without capturing the deeper pedagogical structures that made tutoring effective (Bloom, 1984). Again and again, education mistook polish for thought.

The AI debate reprises this history. Big data, small task promises gains in polish; small data, big task insists that structure is what counts. The first produces graphs that rise, the second produces systems, and students, that endure. The difficulty is psychological as much as technical. Scale feels safe. It generates impressive outputs, marketable products, and measurable progress. Small data feels fragile, risky, slow. But if the purpose of intelligence is to reason in the world as it is, dynamic, unpredictable, changing, then fragility under change is fatal.

Zhu's departure from UCLA dramatised this choice. By walking away from the scale-first mainstream, he was betting on a different map. Not the map of benchmarks and fluency, but the map of structure, causality, and transfer. His wager has implications far beyond AI. It forces education to confront its own temptation: to chase outputs, polish, and fluency, while neglecting the messy work of building understanding that travels. The contrast can be summed up simply. Big data, small task answers the question: How well can you perform when everything is familiar? Small data, big task answers the question: What do you understand when nothing is quite the same? The first produces scores; the second produces intelligence. Schools must decide which question they want to ask.

The Cognitive Science Backbone

It begins with a toddler. She is not quite two, and one afternoon her father points at a neighbour's spaniel and says, "dog." The child has seen dogs before, but not this one: its ears flop differently, its bark is sharper, its coat spotted with brown. Yet she doesn't hesitate. "Dog," she repeats. A week later, in the park, she sees a Dalmatian. Again: "dog." No one has fed her millions of

examples. No cloud server has processed her exposures. A handful of encounters has been enough to build a concept that travels. This scene, so ordinary in a family album, contains a mystery that has preoccupied cognitive scientists for decades. How do humans leap from the sparse to the general? How can a child infer rules from so little? Why does the mind resist the brute-force logic of scale, choosing instead the elegant economy of structure? This was the subject of my own PhD back in the 2000s and remains a fascination to me today. The work of Jerry Fodor, David Chalmers and Dan Dennett et al remains agelessly prescient.

The mystery matters, because the path taken by machines has been the opposite. To produce their fluent answers, large language models are trained on billions of tokens. They consume every imaginable permutation of words and sentences, yet still stumble on tasks that toddlers handle with ease. They confuse cause and effect, falter on counterfactuals, misapply familiar rules in novel contexts. The toddler says “dog” with quiet certainty; the model, even with terabytes of training, might still call a cat a rabbit when the framing is unusual. The explanation lies in what cognitive science has revealed over the past half-century: human learning is not statistical accumulation but structured modelling. We learn not by hoarding data but by building causal, compositional, and social representations that allow us to generalise far beyond what we see.

The story of compositionality is an old one. In 1988, philosophers Jerry Fodor and Zenon Pylyshyn published their famous critique of connectionism, the early neural networks of their day. They argued that human thought is compositional: we can generate and understand new ideas by combining familiar parts in novel ways. The sentence “the red ball rolled under the table” can be understood even if you have never heard it before, because the mind composes known elements (red, ball, rolled, under, table) according to structured rules. Connectionist models of the time struggled with such novelty, and Fodor and Pylyshyn insisted that without compositionality, no system could claim to replicate human thought.

Three decades later, the problem has returned. Modern large language models can generate fluent text, but when asked to compose familiar concepts in unfamiliar ways, they often fail. Lake, Ullman, Tenenbaum, and Gershman (2017) showed that humans can learn new visual concepts from just a handful of examples, recombining strokes and shapes to generate novel instances. Machines trained on thousands of images struggled to do the same. Human cognition, they argued, is fundamentally compositional: it reuses parts to build wholes.

The story of causality runs alongside. Judea Pearl’s landmark work Causality (2000) formalised the distinction between correlation and cause. To predict what will happen when the world changes, one needs not patterns but models of intervention. Children, Alison Gopnik and colleagues showed, are natural causal learners. They perform what look like experiments, dropping spoons, stacking blocks, testing toys, and in the process, infer rules about how the world works (Gopnik et al., 2004). They do not need to see endless repetitions. They need only enough friction to propose and revise models.

This is why a child who sees one or two examples of dogs can generalise to many. She is not memorising pixels or fur patterns. She is inferring structure: legs, tail, bark, role in the world. When new data arrive, the model updates. The leap from sparse to general is not magic; it is causal inference in miniature. Theory of mind adds a third dimension. Premack and Woodruff (1978) asked whether chimpanzees possess it: the ability to attribute beliefs and intentions to others. Humans certainly do. A child watching her mother search for keys understands that the search is guided not by the location of the keys but by her mother’s belief about where they are. This ability to represent mental states, to model not just the world but the minds within it, underpins social learning, communication, and collaboration. Henry Wellman’s decades of research (2014) mapped how theory of mind develops in children, showing how it scaffolds reasoning, empathy, and cultural transmission.

Few-shot generalisation ties these strands together. Joshua Tenenbaum and colleagues (2011) described how humans leap from a handful of data points to structured, generalisable models. Unlike machines, which often require millions of examples, humans infer abstract rules with astonishing efficiency. Elizabeth Carey (2009) showed how children build number concepts through cultural and cognitive scaffolds, learning not by memorisation but by conceptual

integration. Michael Saxe (2019) explored how abstract reasoning emerges from combining innate capacities with social learning. In every case, the theme is the same: humans learn through models, not through brute force. We construct frameworks that are causal, compositional, and social, and those frameworks allow generalisation from the sparse to the vast.

The educational parallels are immediate. Jean Piaget saw children as model-builders, moving through stages of development where they constructed increasingly abstract representations of the world. Lev Vygotsky emphasised the zone of proximal development, where learning occurs through social scaffolding, one mind supporting another to reach just beyond its current grasp. Jerome Bruner spoke of the spiral curriculum, where concepts are revisited at increasing levels of sophistication, each time restructured into richer models. Carey's work on number showed how culture provides the scaffolds for concepts that the brain alone cannot conjure.

Teachers, knowingly or not, enact these theories daily. When a science teacher asks students to predict what will happen to a plant deprived of sunlight, she is asking them to apply causal reasoning. When a literature teacher invites students to compare themes across novels, she is testing compositional transfer. When a maths teacher introduces algebra by showing how variables stand in for quantities, she is scaffolding abstraction. These are not acts of scale; they are acts of structure. Which is why the encounter with scale-first AI has been so unsettling. On the surface, the outputs look impressive. Students can generate essays that mimic fluency, solve equations that mimic competence, write reports that mimic understanding. But under pressure, the brittleness shows. Without structured models, there is no transfer. Without causality, there is no adaptation. Without compositionality, there is no novelty. Teachers sense this instinctively. Cognitive science confirms it empirically.

The revelation, then, is not merely that scale-first AI is brittle. It is that education has always been about what scale cannot provide. The heart of teaching is cultivating transfer under constraint, reasoning under change, understanding that travels. The toddler saying "dog" after two encounters is not a party trick; it is the paradigm of learning. The machine requiring billions of tokens to mimic the same is not evidence of progress; it is evidence of the wrong map.

The Problem of Fluency without Transfer

A literature professor once described her unease after grading a stack of student essays. On the surface, they were extraordinary. The paragraphs were clean, the citations impeccably formatted, the arguments arranged with textbook logic. Yet something was missing. The metaphors felt strained, the interpretations oddly generic, the voice flat. It was as if the students had learned how to perform the act of essay writing without ever grappling with the literature itself. When she pressed them in seminar, the hollowness showed. Asked to compare two characters or reframe an argument under a new lens, they faltered. The polish was there; the transfer was not.

This unease has become familiar to teachers across subjects. In mathematics, a Year 9 student breezes through a worksheet of algebraic equations. Every answer is correct, every step laid out neatly. Yet the same student freezes when faced with a word problem about dividing a recipe into portions or calculating the speed of a train. The knowledge that seemed secure collapses when framed differently. In science, a group of pupils can reproduce the stages of photosynthesis in an exam, but when asked how a drought might affect a food web, their reasoning fragments. The facts are intact; the understanding is brittle.

The arrival of AI has magnified these patterns. Large language models can generate essays indistinguishable from polished student work. They can solve equations, draft lab reports, produce persuasive speeches. To a casual eye, the results look like mastery. But as researchers have shown, the outputs crumble when pushed beyond the grooves in which they were trained. Madaan and colleagues (2023) demonstrated that models able to solve familiar algebraic problems failed on slight variations, misapplying procedures rather than reasoning flexibly. Schaeffer, Miranda, and Koyejo (2023) argued that many so-called emergent abilities are mirages, artefacts of evaluation rather than genuine generalisation. MIT's counterfactual tests revealed brittleness under minimal shifts in framing (MIT CSAIL, 2024).

Gary Marcus has called this phenomenon “fluent nonsense” (Marcus & Davis, 2019). The words flow, the answers impress, but the structure is hollow. The systems excel at fluency without transfer: outputs that look convincing but lack the resilience to travel across contexts. Education has seen this before. In the mid-20th century, programmed instruction promised efficient learning through teaching machines. Students answered questions in linear frames, receiving reinforcement for correct responses. Performance improved, on the tests built to measure those responses. But when asked to apply knowledge in novel contexts, the gains evaporated. The machines had produced fluency, not understanding (Suppes, 1966).

Benjamin Bloom’s famous “2 sigma problem” (1984) offers another cautionary tale. Bloom showed that one-to-one tutoring could lift student performance by two standard deviations: a staggering effect. But when schools tried to replicate the results through programmed instruction or simplified scaffolds, the gains disappeared. The surface features of tutoring could be copied; the adaptive transfer at its core could not.

The same mirage plays out with AI. Fluency feels like progress. Rising scores feel like learning. But without transfer, both are illusions. The psychology of this illusion is powerful. Humans are drawn to surface polish. We equate eloquence with intelligence, neatness with mastery, speed with understanding. In classrooms, this translates into rewarding tidy handwriting, well-structured essays, correct answers. In AI, it translates into celebrating benchmark scores and glowing outputs. In both cases, we are vulnerable to mistaking appearance for substance.

Cognitive science explains why this is dangerous. Sparrow, Liu, and Wegner (2011) described the Google Effect on memory: when information is readily available, people offload recall to external systems. Risko and Gilbert (2016) extended this, coining the term cognitive offloading to describe how we delegate memory, calculation, and problem-solving to tools. Offloading is not inherently bad, writing itself is a form of cognitive offloading, but it carries risks. When offloading replaces reasoning, the underlying capacity atrophies. When students rely on calculators without grasping number sense, or on AI without grasping structure, the result is fluency without depth.

The danger in education is that the offloading is invisible. Teachers see the polished essay, not the brittleness beneath. Inspectors see rising scores, not the lack of transfer. Policymakers see benchmarks ticked off, not the fragility hidden behind them. Just as AI systems fool us with eloquence, students can pass assessments without ever internalising the causal, compositional models that make knowledge travel.

The case studies are stark. In literacy, large language models can draft essays that persuade on the surface. But when asked to adapt metaphors, connect themes across texts, or respond to counterarguments, the structure fails. The machine’s eloquence masks an absence of interpretive depth. The student who submits such work learns performance, not reasoning.

In mathematics, AI tools solve algebraic equations with apparent mastery. But when word problems shift the framing, embedding the same algebra in cooking, travel, or physics, the systems falter. Students who rely on such tools risk absorbing procedures without understanding, competence without flexibility.

In science, AI can produce lab reports that mimic structure: hypothesis, method, results, conclusion. Yet when the teacher alters a variable, changing temperature, introducing new conditions, the reasoning collapses. The report was fluent, but the model behind it was brittle.

These failures mirror the research literature. Schaeffer et al. (2023) on mirage abilities; Marcus on fluent nonsense; Tenenbaum et al. (2011) on the need for structured generalisation; Lake et al. (2017) on compositional learning. Together, they reveal the gap between fluency and transfer. For teachers, this gap is not abstract. It is lived daily in classrooms. A Year 10 student can recite the formula for acceleration but cannot apply it when asked about a cyclist climbing a hill. A Year 7 pupil can list the causes of the English Civil War but cannot reason through how events might have unfolded differently. A Year 12 economics class can memorise models of supply and demand but falters when asked to apply them to contemporary crises. In each case, fluency has been achieved, but transfer has not.

The risk is that AI, if uncritically adopted, entrenches this pattern. Schools already face pressure to "teach to the test." AI tools optimised for benchmarks amplify that logic. Students learn to produce the outputs that look right, while the structures of reasoning remain untested. Assessment rubrics reward polish, procurement contracts demand glowing metrics, policymakers celebrate rising scores. The cycle feeds itself. The alternative is harder but more honest. It means designing tasks that resist offloading—tasks that demand explanation, counterfactual reasoning, adaptation under change. It means building rubrics that reward transfer, not just fluency. It means treating fluency as the starting point, not the end.

The Flip of the Frame

In a Year 6 classroom, a teacher sets her students a peculiar challenge. Instead of asking them to solve maths problems, she asks them to teach the computer. The class is divided into groups. Each group is given a simple agent: an AI program with no prior knowledge of fractions. Their task is to show the agent five examples and then test it. "Here is one-half," they begin, drawing circles divided into equal parts. "Here is one-quarter." Then they switch the framing. "What about three-quarters?" The agent stumbles. The students laugh, then regroup, adding new examples, refining their explanations.

It is a playful exercise, but beneath it lies a radical inversion. The classroom has flipped from a space of answer reproduction to a laboratory of model-building. The students are no longer passive recipients of knowledge or clever prompt engineers coaxing outputs from a fluent system. They are active interrogators, probing the agent, repairing its rules, testing its limits. The point is not to get polished answers but to expose brittle ones and refine them. This inversion captures the shift from prompting to probing. Prompting assumes that fluency is the measure: how well can I phrase a question to elicit a convincing response? Probing assumes that resilience is the measure—how well can I test, adapt, and improve the model when conditions change? Prompting celebrates surface; probing pursues structure.

The logic of prompting has dominated early AI adoption in classrooms. Teachers swap tips on how to write the perfect prompt for an essay plan, a quiz, a lesson sequence. Students learn that a well-crafted prompt yields smoother prose, better summaries, cleaner diagrams. But the very idea of prompt-craft risks entrenching the problem. It trains students to accept outputs as they are, to polish the surface rather than interrogate the depths.

Probing, by contrast, invites scepticism. It treats the machine not as an oracle but as a sparring partner. The goal is not to extract eloquence but to reveal brittleness. A student who probes an AI system does not ask only for an essay on climate change; she asks what would happen if rainfall patterns reversed, if carbon taxes failed, if technology advanced unevenly. She pushes the system to its edge, and in doing so, strengthens her own reasoning.

This shift has deep roots in educational theory. Jerome Bruner argued for discovery learning, in which students construct understanding by exploring and testing rather than memorising. Seymour Papert's constructionism invited children to teach the computer turtle in LOGO how to move, arguing that the act of teaching the machine deepened the child's own understanding. Marlene Scardamalia and Carl Bereiter spoke of classrooms as knowledge-building communities, where students refine shared models rather than consume static facts. In each case, the emphasis was on interrogation, construction, transfer.

The probing frame brings these traditions into the age of AI. Where Papert had the turtle, today's classrooms have language models. Where Bruner spoke of discovery, today's teachers can design assignments that require agents to be tested, challenged, and repaired. The classroom becomes less about what the machine produces and more about how the student engages with its failures.

The implications ripple outward. Assessment, for one, looks different. Imagine a rubric that includes not just "clarity" and "accuracy" but also Explainability and Counterfactual Robustness. A history essay might require not only describing the causes of the French Revolution but probing what would have happened had bread prices stabilised. A science project might involve building a

model of an ecosystem, then testing its resilience under drought or invasive species. A maths exam might include standard equations alongside tasks that demand reframing and adaptation.

This is what probing makes visible: the difference between polish and structure. The student who relies on prompting can produce a fluent essay. The student who probes must wrestle with causality, compositionality, and transfer. One polishes; the other learns. The flip of the frame also demands a cultural shift in how schools think about technology. The first wave of AI adoption has often been framed as efficiency: saving time, producing resources, generating content. But efficiency is a false friend if it entrenches brittleness. A perfectly efficient system that produces hollow understanding is worse than none at all. The second wave, probing, frames AI not as a productivity tool but as a mirror. Its brittleness reflects our own vulnerabilities. Its failures expose what real learning requires. A system that stumbles on counterfactuals reminds students why causal reasoning matters. An agent that misuses metaphors reminds them why interpretation cannot be reduced to pattern recognition. In this sense, AI becomes less a tutor and more a foil, a partner in the Socratic dialogue, whose missteps sharpen human thought.

Consider a Year 10 English class reading Of Mice and Men. A student asks the AI to write an essay on loneliness in the novel. The output is smooth, generic, technically competent. Another student, guided by the teacher, takes a probing approach. She asks the AI how the theme of loneliness would shift if the character of Curley's wife had survived longer in the narrative. The system falters, producing incoherent speculation. But in that failure lies the learning. The student sees that genuine interpretation requires causal and thematic reasoning, not surface patterning. She refines her own analysis, not by accepting the machine's answer but by testing its brittleness. Or take science. A Year 9 group models ecosystems. The AI can list food chains fluently, but when asked to predict the effects of reduced rainfall, it stumbles. The students press further: what if rainfall dropped by 30% for two years? What if new predators arrived? Each probe reveals gaps. The class, in dialogue with the AI, builds resilience into their own models. The machine's weakness becomes the students' strength.

The move from prompting to probing is not a minor adjustment. It flips the frame of the classroom itself. In the prompting frame, AI is a tool for producing answers more quickly. In the probing frame, AI is a tool for revealing the difference between answers and understanding. In the prompting frame, the student is a consumer of fluency. In the probing frame, the student is a critic of brittleness. This is why the flip is so powerful. It turns what looks like a threat, the ease with which machines can generate fluent but hollow work, into an opportunity. It forces schools to redesign pedagogy around what machines cannot do: transfer, causality, explanation. It reframes AI not as the end of learning but as its provocation.

The policy implications are equally profound. Procurement teams must stop asking whether a system can generate content and start asking whether it can support probing. Can the agent survive counterfactual tests? Can it explain its reasoning? Can it adapt under sparse examples? National strategies, whether the UAE's AI 2031 roadmap, the EU AI Act, or U.S. EdTech guidelines, should ask the same. The danger is not that AI fails schools, but that schools fail to ask the right questions of AI. In this sense, Zhu's courtyard paradigm returns. His Tong Test was designed not to reward fluency but to test resilience under dynamic, embodied conditions. The probing classroom is the human analogue. Both reject the surface measures of benchmarks and embrace the deeper test of transfer. Both insist that intelligence is revealed only under change.

In the end, the flip of the frame is about identity. Do we want students to be prompt engineers, adept at coaxing polished outputs from brittle systems? Or do we want them to be model-builders, capable of probing, questioning, and transferring knowledge across contexts? The answer will shape not only how AI is used in schools but what kind of learners those schools produce. The mystery that began with brittle AI systems has led us here: to a new vision of the classroom itself. The revelation is not just that AI must change, but that schools must change too. The path forward is not prompting for fluency but probing for understanding.

Curriculum Transformation

On a rainy morning in Tampines, a suburb of Singapore, thirty secondary students file into a science lab where the teacher has set up a simulation of rainfall across a digital ecosystem. At first, the model behaves predictably: grasses sprout, herbivores graze, predators stalk in tidy cycles. Then the rainfall drops. Within seconds the simulation buckles; plants wilt, prey species collapse, predators starve. The students are not asked to describe what they see. They are asked why. One sketches causal chains on the whiteboard, another suggests testing a new variable: what if rainfall decreases by half, but an invasive plant arrives at the same time? The system is tweaked, the results are surprising, and soon the whiteboard is covered in arrows, loops, and hastily scribbled counterfactuals.

This is not the kind of lesson that can be reduced to multiple-choice tests. It demands that students generate explanations robust enough to survive when the world changes. The principle is familiar in cognitive science, where research on transfer has shown repeatedly that the ability to apply knowledge across contexts is a better predictor of deep learning than rote performance (Perkins and Salomon, 1992; Barnett and Ceci, 2002). Inquiry-based science lessons of this kind echo findings from Hmelo-Silver's studies of problem-based curricula (2007) and from Krajcik and Blumenfeld's work on project-based science (2006), where the emphasis falls on students' capacity to adapt models, not simply recall facts.

Mathematics offers its own lens on the issue. For decades, researchers such as Schoenfeld (1985) and Kilpatrick (2001) have shown that students can display procedural fluency while lacking conceptual understanding. A class may master the steps of solving simultaneous equations, yet when those equations are embedded in a context, say, determining the mixture of two alloys with different properties, the reasoning collapses. One approach that has gained traction is the deliberate design of adversarial or "trapdoor" word problems, where rehearsed procedures are insufficient and students must rely on deeper structures (Boaler, 2016; Star and Newton, 2009). The resonance with Zhu's paradigm is striking: intelligence is revealed not by how many problems one can solve in distribution, but by how one copes when the distribution shifts.

History classrooms have long recognised the same tension. Students can recite the causes of the Cold War in neat bullet points, ideology, spheres of influence, the arms race, but historians like Wineburg (2001) and Seixas and Morton (2013) have shown that real historical thinking requires weighing evidence, considering alternatives, and engaging with counterfactuals. A teacher who asks, "What if the Berlin Wall had never fallen?" is not inviting fantasy but testing whether students have built a causal model robust enough to travel. In the Canadian "Benchmarks of Historical Thinking" project, counterfactual reasoning is explicitly named as a sign of advanced literacy, precisely because it distinguishes surface recall from structured understanding.

Literature, too, depends on transfer. A Year 11 class reading Macbeth may produce essays fluent in theme and quotation. But when asked to consider how the play would change if Lady Macbeth had survived into Act V, the students confront a different order of reasoning. They must recombine themes, test narrative structures, and imagine consequences. The act resembles what Gentner (2001) describes as structural alignment in analogy: the cognitive work of mapping relations rather than reproducing labels. When teachers integrate such probing tasks, they create what Langer (2011) has called "envisionment building," the extension of literary interpretation beyond the given text into new conceptual terrain.

Even in the earliest years of schooling, the principle holds. Gopnik and colleagues (2004) famously described children as "little scientists," testing causal hypotheses through play. A preschooler stacking blocks is conducting miniature experiments: what if the heavy block goes on top, what if the tower leans, what if the foundation shifts? Siegler's overlapping waves theory (1996) showed that children flexibly switch strategies under new conditions, revealing not rehearsed fluency but adaptive reasoning. Early years curricula in Finland and Reggio Emilia settings in Italy have long incorporated such exploratory play, and longitudinal research indicates that it fosters long-term problem-solving ability (Broström, 2017).

What emerges across these domains is a curriculum that looks less like a ladder of rehearsed skills and more like a field of probes. Knowledge is not treated as static content to be recalled, but

as models to be tested under strain. The humanities invite students to reimagine causes and outcomes, literature demands recombination of themes, mathematics requires resilience under shifting contexts, science insists on causal diagrams robust to variable change, early years nurture experimentation in play.

The difficulty, of course, is that many curricula remain yoked to assessments that prize fluency. Black and Wiliam's work on formative assessment (1998) warned that systems oriented around narrow tests encourage shallow learning, and more recent studies have shown that high-stakes regimes can diminish transfer (Au, 2007). The OECD's reports on global competencies (2018, 2021) argue that adaptability and problem-solving should be central, but national curricula often fall back to content coverage and replicable performance. This is why the paradigm shift is so urgent: once machines can produce fluency more reliably than students, a curriculum built on surface outputs is untenable.

Some countries have begun to move. Singapore's science curriculum now incorporates "science practical assessments" where students must design and explain investigations under novel conditions. Finland's "phenomenon-based learning" approach encourages students to integrate multiple disciplines in projects that resist rote answers (Lonka, 2018). The UAE has piloted AI literacy modules that focus less on tool use and more on critical interrogation of outputs (KHDA, 2023). Each of these reforms hints at a curriculum already edging toward small-data, big-task logic.

But the transformation cannot remain piecemeal. Zhu's paradigm implies a systemic redesign: subjects structured around opportunities for probing, tasks constructed to reveal transfer, assessments aligned to reward resilience. This is not an argument against content knowledge, decades of research from Hirsch (1987) to Kirschner, Sweller, and Clark (2006) confirm its necessity. It is an argument that knowledge must be taught and tested in ways that expose its structure, not merely its surface.

The stakes are high. Curricula that remain bound to fluency risk producing students who look competent but falter under novelty, a mirror of the brittle machines they increasingly rely on. Curricula that embrace probing may cultivate learners who, like the Singapore students drawing arrows on the whiteboard, see knowledge not as fixed but as a system of models constantly under revision. That difference, between brittle fluency and resilient transfer, may decide whether schools prepare students for a world of changing conditions or leave them rehearsed for tests that no longer matter.

Assessment Revolution

In 2017, Singapore's Ministry of Education announced that the long-standing science practical examination at the secondary level would be restructured. Instead of testing whether students could reproduce rehearsed experiments, the assessment would ask them to design and explain investigations under unfamiliar conditions. The shift, small on the surface, was radical in intent. Students would no longer be rewarded merely for executing procedures they had practised; they would be judged on their ability to adapt their reasoning when variables were altered. The ministry explained that this was meant to reflect "the authentic practices of scientists" and to cultivate transfer of knowledge beyond the laboratory bench (Singapore Ministry of Education, 2018).

This reform exemplifies the broader problem and opportunity of assessment in the age of AI. For decades, most examinations have prized fluency: polished essays, tidy problem sets, correct answers in familiar frames. But as Black and Wiliam's seminal review of formative assessment argued, performance on such tasks tells us little about underlying understanding (1998). The cognitive revolution in assessment research, consolidated in Pellegrino, Chudowsky, and Glaser's Knowing What Students Know (2001), insisted that tests must capture the mental models students hold, not just the outputs they can reproduce. Yet in practice, high-stakes systems continued to default to fluency, partly because it was easy to measure and easy to rank.

AI has broken that fragile bargain. Large language models can generate essays that mimic fluency, but when pressed into counterfactual or explanatory reasoning, they reveal brittleness

(Marcus and Davis, 2019; Schaeffer et al., 2023). In mathematics, models can solve algebraic equations but fail when problems are reframed into novel contexts (Madaan et al., 2023). In science, they can produce polished lab reports that collapse when conditions shift (MIT CSAIL, 2024). If machines can now produce the outputs we once used to judge students, then those outputs can no longer serve as reliable indicators of learning.

Some systems have recognised the urgency. Finland's curriculum reform of 2016, which embedded phenomenon-based learning projects, redefined assessment to include interdisciplinary tasks requiring transfer across domains (Lonka, 2018). Students might investigate climate change through data analysis, historical interpretation, and ethical debate, with performance judged on the ability to integrate and adapt. Early evaluations suggest that while such assessments are more complex to administer, they provide a richer picture of student reasoning (Krokfors et al., 2021). In Dubai, the Knowledge and Human Development Authority (KHDA) has piloted "innovation strands" in inspections that look not only at exam performance but at how schools foster problem-solving and creativity, requiring evidence from portfolios, oral defences, and student projects (KHDA, 2023). These shifts are uneven, but they reveal a recognition that traditional examinations risk obsolescence.

The central task is to design assessments that expose structures of thought rather than surfaces of performance. In mathematics, this might mean asking students to explain why a solution holds when numbers change, echoing research that shows conceptual explanation correlates more strongly with long-term mastery than procedural repetition (Boaler, 2016; Star and Newton, 2009). In science, it might involve giving students incomplete data sets and asking them to infer causal mechanisms, an approach Lehrer and Schauble (2006) found effective in fostering model-based reasoning. In history, it could mean embedding counterfactual tasks: what if bread prices had stabilised in late eighteenth-century France? Would revolution have followed? Wineburg (2001) and Seixas and Morton (2013) argue that such reasoning distinguishes novice recall from expert historical thinking.

Rubrics would need to evolve accordingly. Instead of grading only clarity and accuracy, they could include categories such as Explainability, Counterfactual Robustness, and Transferability. These dimensions align with Pellegrino and Hilton's framework in Education for Life and Work (2012), which argued that adaptability and reasoning across contexts are essential twenty-first-century competencies. They also reflect the OECD's Learning Compass 2030 (2019), which positions student agency, resilience, and problem-solving at the core of future-ready education. UNESCO's 2023 guidelines on AI in schools highlight a similar principle: that assessment must capture what cannot be outsourced to machines.

The shift also requires cultural change. Parents and policymakers often equate rising test scores with progress, but history shows how misleading this can be. Bloom's "2 sigma problem" demonstrated that one-to-one tutoring dramatically raised performance, but when tutoring was reduced to programmed instruction, the gains disappeared (Bloom, 1984). The lesson was that adaptive feedback, not drill, drove success. Likewise, assessments that reward rehearsed fluency will generate rehearsed learning, whether by human students or their AI assistants. Only when assessments reward transfer will curricula adapt to cultivate it.

The ethical stakes are clear as well. If assessment systems continue to reward fluency, students with access to advanced AI tools will gain an advantage that reflects not their reasoning but their resources. By contrast, assessments that demand explanation and resilience are harder to outsource. They level the field by privileging reasoning over reproduction. In this sense, rethinking assessment is not only a pedagogical imperative but an equity one.

The reforms underway in Singapore, Finland, and Dubai are still partial, but they point toward a new logic. Assessment must become less about what students can polish and more about what their knowledge can survive. That requires designing tasks where offloading is insufficient, where brittleness is exposed, and where understanding must travel. In the end, what matters is not how well a student, or a machine, can reproduce answers under familiar conditions, but how they reason when the conditions change.

AI Literacy Redefined

In 2023 UNESCO released its Guidance for Generative AI in Education and Research. Buried within the report was a telling warning: schools rushing to integrate AI risked narrowing "AI literacy" to the mechanical act of operating tools rather than cultivating the deeper capacities needed to interrogate and critique them (UNESCO, 2023). The language was careful but clear. If literacy was reduced to "prompting for performance," students might master a craft that was already on the verge of obsolescence.

The idea of AI literacy has been circulating for some years, often framed in terms of functional skills. Curricular pilots in the United States and Europe have encouraged teachers to introduce "prompt engineering" into classrooms, positioning it as a twenty-first century equivalent of touch typing. Early studies suggest that students can indeed learn to manipulate phrasing in order to improve the outputs of large language models (Kasneci et al., 2023). Yet as with the first wave of "digital literacy" programmes two decades ago, there is a danger that operational competence is mistaken for intellectual capacity. Gilster's definition of digital literacy in the 1990s insisted that the real skill lay in critical evaluation and cross-referencing, not simply in clicking or typing (Gilster, 1997). The same distinction now applies to AI: using it is not the same as understanding it.

Prompt engineering promises fluency but does not cultivate transfer. A well-phrased query can generate a persuasive essay, but it does not train a student to weigh sources, question assumptions, or revise reasoning under changing conditions. Marcus and Davis (2019) showed that large language models remain brittle under even modest shifts in framing, producing outputs that are eloquent but logically inconsistent. Choi (2023) described commonsense as the "dark matter of intelligence," largely absent from the statistical patterns on which such systems rely. Teaching students only to polish prompts risks leaving them vulnerable to these weaknesses, mistaking surface eloquence for depth.

Probing represents an alternative frame. Instead of rewarding students for eliciting the smoothest outputs, it trains them to test the edges of the system: to ask what happens when variables are reversed, when counterfactuals are introduced, when causal chains must be explained. In this sense, probing aligns with traditions of discovery and construction that long pre-date AI. Papert's constructionism in the 1980s urged children to "teach the turtle" in LOGO, reasoning through the act of debugging a machine (Papert, 1980). Scardamalia and Bereiter (2006) framed classrooms as knowledge-building communities where the emphasis lay not on reproducing information but on refining models. The spirit of these approaches lies in interrogation, not extraction.

There are already attempts to codify this. The OECD's Learning Compass 2030 situates agency, resilience, and transfer as the core competencies of future readiness (OECD, 2019). The European Commission's DigComp 2.2 framework includes critical interaction with AI outputs as a marker of digital competence (Vuorikari et al., 2022). In the UAE, KHDA's guidance on AI in education stresses that literacy should mean recognising bias, testing validity, and understanding limitations rather than perfecting queries (KHDA, 2023). Each of these documents gestures toward a richer conception of literacy than prompt-craft.

Educational psychology provides the grounding for why this matters. Vygotsky's concept of the zone of proximal development illustrates that learning occurs when students are supported to stretch just beyond their current capabilities (Vygotsky, 1978). If AI literacy is defined as prompt efficiency, the zone narrows to optimisation. If it is framed as probing, the zone expands to include the reasoning gaps exposed by machine brittleness. Bruner's notion of a spiral curriculum likewise emphasises revisiting concepts at increasing levels of abstraction (Bruner, 1960). Applied to AI, this means moving beyond initial exposure to tool use toward sustained engagement with how models fail and why.

Recent empirical studies support this orientation. Greenhow and Askari (2021) found that students working with AI-based writing assistants improved their surface correctness but showed no gains in critical reasoning. In contrast, Luckin et al. (2022) argued that AI systems used as objects of interrogation, where students analysed and critiqued outputs, did foster metacognitive

awareness. Chiu and Lim's (2023) work with Singaporean secondary students demonstrated that explicit discussions of AI limitations encouraged deeper understanding of scientific models, suggesting that probing activities can function as scaffolds for conceptual growth.

The stakes are pedagogical and political. If AI literacy is equated with prompt-craft, schools risk producing a generation trained to interface smoothly with systems they cannot question. This echoes the early computer literacy programmes of the 1980s that focused on learning BASIC syntax rather than cultivating computational thinking (Wing, 2006). Within a decade, the operational skill was obsolete, while the underlying reasoning remained scarce. The same pattern is poised to repeat unless literacy is defined as critique rather than operation.

The alternative is to imagine AI literacy as a mirror for human reasoning. When a system produces a fluent but flawed answer, the student's task is not to refine the prompt but to diagnose the error. When an essay reads smoothly but fails under counterfactual challenge, the student must articulate why. When a model misapplies metaphor, the student can learn what genuine analogy requires. In each case, the literacy lies not in coaxing fluency but in recognising brittleness, and then building understanding that surpasses it.

The frameworks are beginning to shift. UNESCO, OECD, the EU, and national agencies are all edging toward definitions of literacy that prioritise probing over prompting. But curricula remain tempted by the simplicity of teaching operational skills. The tension is unresolved, and schools stand at the fork. They can define literacy as the knack of getting a machine to produce polished text, or they can define it as the ability to interrogate, test, and refine models. One definition risks aligning students with brittle systems. The other aligns them with the deeper structures of human learning. The choice is not trivial. It will shape how the next generation relates to intelligence itself.

Equity and Ethics

In 2021, a group of African researchers published a short paper that circulated quietly but struck a chord among those who read it. Titled The Elephant in the Room: Large Language Models in Africa, it argued that the costs of training and running frontier AI models placed them entirely out of reach for most universities and schools on the continent (Nekoto et al., 2021). While Silicon Valley companies boasted about scaling billions of parameters, the authors pointed out that many African research centres struggled with intermittent electricity, limited bandwidth, and compute budgets that could barely support modest experiments. The gap was not merely technical but moral: who gets to define the trajectory of intelligence when only a handful of well-resourced players can participate?

This question lands heavily in education. If AI tools require massive compute resources and continuous data extraction, then access will inevitably be unequal. Students in wealthy systems may interact daily with advanced tutors; those in low-income contexts may be excluded altogether. Bender et al. (2021) warned of the ecological costs of "stochastic parrots," calculating the carbon footprint of training large language models as equivalent to the lifetime emissions of several cars. Strubell et al. (2019) estimated that the energy used in training a single large model could power multiple households for a year. When education systems adopt such tools without scrutiny, they risk embedding inequities of both access and environmental burden.

Small-data, big-task approaches offer a different horizon. Models that learn effectively from sparse examples can be run locally on modest hardware, reducing both cost and surveillance. Xu et al. (2023) demonstrated that smaller causal-reasoning models could perform competitively on transfer tasks without requiring vast corpora. This matters for schools. A rural classroom in Kenya or Nepal is far more likely to benefit from a lightweight, transparent agent that can run offline than from a massive proprietary model that demands high-speed connections and constant cloud access. When Zhu speaks of building "general intelligence with small data," the ethical implications extend beyond pedagogy: they sketch a path for inclusion.

Privacy is bound up in the same story. Big-data models depend on collecting and processing enormous amounts of information, much of it from users themselves. In education, this often translates into harvesting student work, classroom interactions, even keystroke patterns. Selwyn

(2022) has argued that the "datafication of education" risks normalising surveillance, with children's intellectual development becoming raw material for optimisation. By contrast, small-data models reduce the pressure to hoard. If systems can learn from limited, curated examples, the rationale for indiscriminate collection diminishes. The ethical upside is not just efficiency but dignity: students can learn without being mined.

There is also a cultural dimension. Large-scale models trained predominantly on English-language data reproduce biases and marginalise local knowledge systems (Joshi et al., 2020). In education, this means that AI tutors may reflect perspectives far removed from students' contexts. Smaller, localised models, built with community data and transparent structures, allow for greater cultural relevance. Sabelo Mhlambi (2020) has argued for "Decolonial AI," insisting that systems must reflect the epistemologies of the communities they serve. In a classroom, that could mean an agent that recognises local proverbs, historical narratives, or ecological practices, not one that imposes generic global content.

Equity in education has always been tethered to access to tools, from textbooks to broadband. The danger with AI is that the tools themselves may accelerate inequality if their cost and scale place them in the hands of a few. Policy documents are beginning to register this. The OECD's AI and the Future of Skills report (2021) warned that "without attention to infrastructure and localisation, AI in education risks amplifying existing divides." UNESCO's 2023 guidance emphasised that equitable AI requires "contextualisation and proportionality," not simply importing frontier models into classrooms. The EU's AI Act draft classifies education as a high-risk domain, demanding safeguards that include transparency and fairness (European Commission, 2023).

The environmental costs compound the ethical picture. With education systems under pressure to demonstrate sustainability, deploying AI tools that burn vast amounts of energy is increasingly untenable. Luccioni et al. (2022) developed methodologies for estimating the carbon footprint of training and inference, showing that the marginal costs of scaling often outweigh the benefits. For schools and ministries committed to climate goals, adopting lightweight, explainable systems is not only a matter of equity but of alignment with broader ethical commitments.

None of this implies that AI should be rejected in education. Rather, it insists that the paradigm chosen matters. Big-data, small-task models risk centralisation, surveillance, inequity, and ecological harm. Small-data, big-task models promise localisation, privacy, inclusion, and sustainability. The choice is not neutral. It is a choice about whose knowledge counts, who has access, and how much the planet can bear.

The Kenyan researchers who described the elephant in the room were not merely critiquing access; they were pointing to the future of fairness in intelligence itself. For schools, the same elephant now stands in the classroom. The ethics of AI are not abstractions: they will shape whether students are treated as data points or as learners, whether knowledge is imported or built, whether intelligence is centralised or shared.

Procurement & Policy

When the European Union released the draft of its AI Act in 2021, education was listed alongside healthcare and policing as a "high-risk domain" (European Commission, 2021). The classification meant that any AI system deployed in schools would need to meet strict requirements for transparency, accountability, and fairness. At first glance the policy seemed distant from classrooms, another piece of Brussels legislation. But for ministries of education and school operators, it was a signal that procurement could no longer be treated as a matter of buying shiny new tools. It was now a question of ethics and governance.

Procurement has always shaped educational technology. In the early 2000s, the race to install interactive whiteboards across the UK consumed millions of pounds with little evidence of impact (Moss et al., 2007). More recently, one-to-one laptop initiatives in the United States revealed how quickly enthusiasm could outpace planning, leaving schools with expensive devices but no change in pedagogy (Zheng et al., 2016). The pattern risks repeating with AI. Vendors arrive with

glossy brochures boasting of parameter counts, benchmark scores, and "state-of-the-art" models. Ministries are told that bigger means better, that scale guarantees performance. But as Zhu's critique shows, scale does not equate to understanding, and fluency does not ensure transfer. The wrong procurement decisions could lock schools into brittle systems that dazzle in demonstrations but falter in practice.

Some governments have begun to ask different questions. The UAE's National Artificial Intelligence Strategy 2031 places education as one of its central pillars. The strategy commits to "embedding AI in teaching and learning" while emphasising localisation and capacity-building (UAE Government, 2019). In practice this means that procurement discussions are not only about technical specifications but about alignment with national goals, cultural relevance, and sustainability. A vendor offering a large, cloud-based model may impress with capabilities, but a smaller, locally adaptable system that can run with limited data may align more closely with both policy and pedagogy.

International organisations have attempted to provide frameworks. UNESCO's Guidance for Generative AI in Education and Research (2023) urges ministries to evaluate systems against criteria such as transparency of training data, explainability of outputs, and proportionality of data collection. The OECD's AI and the Future of Skills report (2021) suggests that procurement officers ask whether tools enable transfer, foster resilience, and reduce inequities. These frameworks are not yet binding, but they illustrate a shift in language: away from size and speed, toward explainability and adaptability.

For schools themselves, the implications are immediate. A headteacher in Helsinki choosing between two AI writing assistants may be presented with accuracy scores and fluency ratings. But the more pertinent questions are different: Can the system explain its reasoning when challenged? Can it adapt from a handful of examples? Can it run locally without extracting student data to remote servers? Research on explainable AI in education has shown that students learn more effectively when systems make their reasoning transparent (Holstein et al., 2019; Khosravi et al., 2022). Yet such features are rarely highlighted in procurement pitches.

The problem is compounded by accountability pressures. Governments often demand evidence of "impact," which vendors provide in the form of benchmark results. But benchmarks can mislead. Schaeffer et al. (2023) showed that apparent emergent abilities in large models often collapse under counterfactual testing. Madaan et al. (2023) found that algebra performance could mask brittleness under reframing. A procurement process that rewards benchmark performance risks enshrining fluency without transfer as the standard.

The alternative is to embed Zhu's small-data, big-task criteria directly into procurement. Imagine a checklist: Can the system learn effectively from limited examples? Does it withstand counterfactual probes? Does it reduce reliance on mass data harvesting? Can it be audited for bias and explainability? These are not abstract aspirations; they are practical filters that align purchasing with pedagogy. They resonate with the EU AI Act's requirements for transparency and human oversight, with UNESCO's calls for proportionality, and with national strategies that tie AI to local values.

Some ministries have begun to adapt their tendering processes. Reports from Singapore indicate that vendors are now required to demonstrate not just functionality but alignment with inquiry-based pedagogical goals (Singapore Ministry of Education, 2022). In the UAE, AI pilots in schools are evaluated partly on cultural alignment and data governance, not only on performance (KHDA, 2023). In Finland, phenomenon-based projects are assessed for their contribution to broader competencies before being scaled nationally (Lonka, 2018). These are early steps, but they show how procurement can be steered toward systems that support probing and resilience.

The stakes extend beyond pedagogy to sovereignty. If education systems rely exclusively on large, foreign-built models, they risk ceding control of both curriculum and data. The debate echoes earlier struggles over textbooks and curricula supplied by multinational companies, now amplified by the scale of AI. Mhlambi's call for "decolonial AI" (2020) highlights how procurement choices determine whose epistemologies shape learning. A school that adopts a locally trained,

explainable model may preserve cultural nuance; one that outsources entirely may embed alien assumptions in daily lessons.

Procurement may sound like bureaucracy, but in the age of AI it is the hinge of ethics and pedagogy. The decision to buy one system rather than another is also a decision about what kind of intelligence a school values. If the metric remains fluency, then contracts will reward the largest, glossiest systems. If the metric shifts to transfer, explainability, and equity, then procurement can become a lever for aligning education with the structures of human learning.

The EU's classification of education as a high-risk domain was not hyperbole. It was recognition that choices made in ministry offices and school board meetings will shape how a generation encounters intelligence itself. The question is not only what systems can do, but what they encourage us to become.

If We Stick with Big Data, Small Task

When New York City first experimented with automated essay scoring in the early 2010s, the promise was efficiency. Thousands of scripts could be marked in minutes, saving teachers hours of drudgery. Yet researchers soon found that systems rewarded length and surface features over argument and originality (Perelman, 2014). A decade later, the technology is slicker, the prose more fluent, the scale vastly greater. But the logic is unchanged: systems trained on vast corpora to mimic fluency, unable to test whether reasoning travels. If schools accept this as literacy, then the exam hall becomes a hall of mirrors, students rewarded for producing work indistinguishable from machines already known to be brittle.

The dangers are not only pedagogical but cognitive. Sparrow's study on the "Google Effect" showed how people offload memory when information is easily retrievable online (Sparrow et al., 2011). Risko and Gilbert (2016) extended the concept to show that humans adjust their cognitive strategies when tools are available, relying on external systems rather than internal reasoning. With AI, the risk multiplies. If every essay, lab report, and problem set can be produced by a system trained on terabytes of examples, students may learn that intellectual effort is unnecessary. The immediate result is polished work; the long-term consequence is atrophied reasoning.

Equity fractures deepen in parallel. Wealthier schools and systems purchase access to frontier models with premium features, while under-resourced schools are left with restricted versions. Studies of EdTech adoption have already shown how access gaps magnify inequalities rather than close them (Selwyn, 2016; Bulman and Fairlie, 2016). In the AI era, the divide may be sharper. Those with resources will learn to use AI to accelerate knowledge production; those without will be confined to brittle tools that reinforce surface skills. The promise of democratisation risks becoming a mechanism of stratification.

Privacy is the next casualty. Large-scale AI systems thrive on data. In education, this often means harvesting student work, keystrokes, and classroom interactions. Williamson and Hogan (2020) have documented how platformisation of schooling brings with it new regimes of surveillance, in which children's data are commodified for optimisation. Selwyn (2022) warns that education risks normalising constant monitoring under the guise of personalisation. Big-data, small-task models demand constant feeding; their hunger for information ensures that students' intellectual development becomes raw material. In such a system, privacy is not a right but a cost, traded away for access to tools.

The environmental implications are rarely mentioned in procurement documents but loom large. Strubell et al. (2019) estimated that training a single large language model could emit as much carbon as the lifetime emissions of several cars. Bender et al. (2021) calculated similar costs, warning of the ecological absurdity of scaling without constraint. Luccioni et al. (2022) have since refined methods for measuring carbon footprints of training and inference, finding that the marginal returns of scale diminish even as energy costs rise. For schools and ministries that proclaim commitments to sustainability, adopting such systems without scrutiny risks a profound

contradiction: teaching climate science in classrooms powered by tools accelerating climate harm.

There is also the question of epistemic authority. Large-scale models are built predominantly on English-language internet corpora, reproducing biases and marginalising alternative knowledge systems (Joshi et al., 2020). In classrooms, this means that AI tutors may deliver "universal" explanations that flatten cultural nuance and silence local epistemologies. The problem is not only technical but cultural: whose knowledge counts when education outsources reasoning to systems trained elsewhere? Mhlambi (2020) has argued for "decolonial AI," yet procurement pipelines continue to favour models built in Silicon Valley, embedding one epistemic frame at global scale.

What makes these dangers insidious is that they often appear as progress. Essays look more polished, test scores rise, efficiency increases. The McKinsey reports on education reform in the 2000s promised measurable gains through standardisation and accountability, yet critics noted how such approaches narrowed the curriculum and encouraged teaching to the test (Fullan, 2011). The same trap now looms larger. When fluency is easy to produce, it will be fluency that is measured, and when fluency is measured, it will be fluency that is taught. The cycle reinforces itself until education becomes indistinguishable from the brittle systems it has embraced.

It is tempting to imagine that schools will notice, that teachers will resist. Some surely will. But policy incentives have a gravitational pull. If exams remain designed around polished essays, tidy procedures, and benchmark fluency, then schools will be forced to align. Ministries will procure the systems that perform best on those metrics. Parents will demand tools that produce immediate gains. The incentives of the system will converge on the very paradigm that Zhu warns against. And by the time its brittleness becomes undeniable, a generation of students may already have internalised its limits.

This is the future we might miss: one where classrooms no longer cultivate reasoning under change but reward polish under repetition. One where student agency gives way to cognitive offloading. One where equity fractures widen, privacy erodes, and sustainability commitments are betrayed. One where the appearance of progress conceals the reality of atrophy. If education continues down the path of big data, small task, the risk is not only that machines will become brittle tutors, but that schools themselves will become brittle institutions.

If We Embrace Small Data, Big Task

In 2016 Finland became the first country to embed "phenomenon-based learning" into its national curriculum. Instead of teaching subjects in isolation, schools were encouraged to design projects that required students to integrate knowledge across disciplines. A climate change unit, for example, might combine data analysis, historical interpretation, ethical debate, and local fieldwork. Assessments shifted from rehearsed content to applied reasoning, judged on the ability to adapt models under changing conditions (Lonka, 2018). Early evaluations found that students demonstrated greater transfer of learning and stronger engagement, even if the projects were harder to administer (Krokfors et al., 2021). The Finnish experiment showed that when tasks are designed to test resilience, students rise to the challenge.

This spirit of curriculum as probing rather than recall is visible elsewhere. Singapore's restructuring of science practical assessments, introduced in 2017, moved beyond rehearsed experiments toward open investigations where students must design, justify, and revise methods (Singapore MOE, 2018). In the UAE, KHDA's guidance on AI literacy emphasises not prompt-craft but critique: students are encouraged to interrogate outputs, identify bias, and test validity (KHDA, 2023). Each of these reforms signals an alignment with the small-data, big-task principle: giving learners fewer rehearsed problems but richer opportunities to demonstrate reasoning under change.

The pedagogical consequences are significant. When students are asked to explain why an equation holds across contexts, they begin to internalise structure rather than memorise procedure. When they model ecosystems under drought or imagine alternative outcomes in history, they practise causal reasoning. When they critique AI outputs, they learn to probe

brittleness and refine models. These activities map directly onto the cognitive science findings that human learning is causal, compositional, and social (Tenenbaum et al., 2011; Gopnik et al., 2004; Carey, 2009). They align with Vygotsky's emphasis on scaffolding reasoning beyond the immediate task (1978) and Bruner's vision of a spiral curriculum that revisits concepts at increasing levels of abstraction (1960).

The equity implications are no less striking. Small-data approaches can run on lightweight systems, reducing reliance on high-cost infrastructure. Xu et al. (2023) demonstrated that causal reasoning models trained on limited examples performed competitively on transfer tasks, suggesting that schools without access to frontier hardware could still deploy meaningful tools. In contexts where bandwidth is scarce or electricity unreliable, such systems are far more viable than massive cloud-based models. They also reduce the pressure to extract student data at scale, protecting privacy in ways that large systems cannot (Selwyn, 2022). For countries where surveillance is already a concern, small-data AI offers not only affordability but dignity.

Culturally, locally adaptable models allow education systems to preserve epistemic diversity. Joshi et al. (2020) highlighted how current language models marginalise under-resourced languages, limiting their utility for many communities. By contrast, smaller, domain-specific systems can be trained on local data, incorporating indigenous knowledge and contextually relevant content. Mhlambi's call for "decolonial AI" (2020) finds practical expression here: classrooms where students engage with systems that recognise their culture rather than erase it. In such spaces, AI becomes a partner in sustaining local knowledge rather than a vector of homogenisation.

The ethical benefits extend to sustainability. Training frontier models consumes vast energy resources (Strubell et al., 2019; Bender et al., 2021). Smaller systems trained on curated, sparse data require a fraction of that cost. Luccioni et al. (2022) showed that inference on compact models can be several orders of magnitude less carbon-intensive. For education systems that claim alignment with climate goals, adopting lightweight models is not only practical but morally consistent. The paradox of teaching environmental science with tools that accelerate emissions disappears when the paradigm shifts.

Examples of alignment are already visible. In Singapore, inquiry-based tasks now serve as gateways for students to demonstrate reasoning rather than recall. In Finland, phenomenon projects cultivate transfer across domains. In Dubai, AI literacy pilots require students to probe outputs rather than prompt for polish. These initiatives are not framed explicitly in Zhu's terms, but they embody his principle: less data, richer tasks. They suggest that schools need not wait for frontier labs to solve general intelligence; they can already pivot pedagogy toward the capacities that matter.

If adopted more widely, the consequences could be profound. Classrooms would cease to resemble factories of polished reproduction and become laboratories of reasoning. Students would be trained not as consumers of fluency but as builders of models. Teachers would shift from delivering content to orchestrating probes. Assessments would reward adaptability, not rote. AI would serve not as a shortcut to surface competence but as a mirror that reveals the boundaries of understanding.

The choice is stark but hopeful. Where the big-data paradigm hollows out learning, the small-data paradigm deepens it. Where one entrenches inequity, the other offers inclusion. Where one exhausts resources, the other conserves them. Where one erases culture, the other sustains it. If education embraces small data and big tasks, it may not only resist the brittleness of machines but rediscover its own purpose: cultivating human reasoning that endures when the world changes.

From Prompting to Probing

In the spring of 2022, Singapore's Ministry of Education piloted a revised science assessment in several secondary schools. Instead of repeating familiar experiments, students were presented with an unfamiliar ecological simulation. They were told to explore what would happen if rainfall patterns shifted and to explain the causal mechanisms behind the changes. The task did not

reward rehearsed procedures; it demanded reasoning under change. Observers noted how students moved from initial confusion into animated discussion, sketching models, testing hypotheses, revising explanations (Singapore MOE, 2022). It was a modest reform in a single subject, but it captured a possibility far larger than the classroom: what if education as a whole reoriented itself around this principle?

The proposal is simple in outline, radical in effect. It asks schools to move from prompting to probing. In curriculum, this means designing units where the goal is not recall but representation, where knowledge is treated as a model to be tested rather than a fact to be repeated. In mathematics this might be adversarial word problems that shift conditions; in literature, interpretive tasks that stretch themes across counterfactuals; in science, model-based investigations that resist offloading. Decades of research, from Bruner's spiral curriculum (1960) to Tenenbaum's work on causal learning (2011), point to the same conclusion: resilient understanding emerges when students are asked to construct and adapt, not simply reproduce.

In assessment, the shift is even more consequential. Rubrics must capture transfer, explainability, and counterfactual robustness. This does not mean abandoning clarity or accuracy, but rebalancing them. A student who can solve an equation but cannot explain why it holds under new conditions has not demonstrated mastery. A student who can write a polished essay but cannot adapt an argument when a variable changes has not shown understanding. Pellegrino and Hilton's Education for Life and Work (2012) framed adaptability as the defining skill of the century; assessments that ignore it risk becoming relics.

In literacy, the proposal is to redefine what it means to be "AI ready." Teaching prompt engineering as a form of literacy is the equivalent of teaching BASIC syntax in the 1980s: a narrow operational skill soon to be obsolete. The real literacy lies in probing—recognising brittleness, testing claims, refining models. This aligns with Papert's constructionism, where children learned most by debugging the turtle, and with Bereiter and Scardamalia's vision of classrooms as knowledge-building communities. It also resonates with contemporary warnings from UNESCO (2023) and the OECD (2019) that literacy must mean critical engagement, not mechanical operation.

In policy, the shift requires procurement frameworks that privilege proportionality over scale. Ministries should ask: Can the system learn from few examples? Can it explain its reasoning? Can it run without extracting vast amounts of student data? These questions echo the EU AI Act's designation of education as high-risk, UNESCO's emphasis on proportionality, and the UAE's strategy of localisation in AI 2031. They represent a refusal to be dazzled by parameter counts or benchmark scores. Procurement is pedagogy by other means; what governments buy is what schools will teach.

The contrast between the two futures could not be sharper. In one, schools adopt scale-first systems that reward fluency, and students learn to mimic brittle machines. In the other, they adopt probing systems that reward transfer, and students learn to reason beyond machines. One hollowing, the other renewing.

The Singapore pilot with rainfall is only one example, but it is a fitting image to close on. Students confronted with a simulation that defies rehearsal must think differently. They sketch diagrams, debate variables, revise their models. In that moment, the classroom becomes a laboratory of intelligence itself—not only of the students, but of the systems they are learning to interrogate. The task is small, the data sparse, but the reasoning is rich.

That is the future the proposal points toward: classrooms where the measure of learning is not how well students or machines can produce polished answers under familiar conditions, but how both can reason when the conditions change.

About the Author

Dr. Neil Hopkin is a globally recognised thought leader in international K-12 education, and serves as the Director of Education at Fortes Education.

His extensive academic background includes advising UK government bodies and spearheading significant educational initiatives, particularly with the EdTech, Early Years, Higher Education and Teacher Professional Development fields, equipping him with invaluable insights and expertise.
As the head of Fortes' Academic Leadership Team, Dr. Hopkin is responsible for overseeing academic performance, operational efficiency, curriculum development, and staff professional development across Fortes Education institutions.

FORTES EDUCATION
1981 Enlighten • Empower • Excel

For more information contact Dr Neil Hopkin at:

www.sunmarke.com www.risdubai.com

Bibliography

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. Educational Researcher, 36(5), pp.258–267.

Barnett, S.M. & Ceci, S.J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. Psychological Bulletin, 128(4), pp.612–637.

Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp.610–623.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education, 5(1), pp.7–74.

Bloom, B.S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. Educational Researcher, 13(6), pp.4–16.

Boaler, J. (2016). Mathematical Mindsets: Unleashing Students' Potential through Creative Math, Inspiring Messages and Innovative Teaching. San Francisco: Jossey-Bass.

Bruner, J.S. (1960). The Process of Education. Cambridge, MA: Harvard University Press.

Bulman, G. & Fairlie, R.W. (2016). Technology and Education: Computers, Software, and the Internet. In: E.A. Hanushek, S. Machin & L. Woessmann (eds). Handbook of the Economics of Education, Vol. 5. Amsterdam: Elsevier, pp.239–280.

Carey, S. (2009). The Origin of Concepts. Oxford: Oxford University Press.

Chiu, T.K.F. & Lim, C.P. (2023). Exploring students' critical thinking development through interrogating AI outputs. Computers & Education, 195, 104687.

Choi, Y. (2023). The Dawn of AI Commonsense: Dark Matter of Intelligence. Keynote, ACL 2023.

Christiansen, B. & Rump, C. (2008). Oral examinations in mathematics: Effects of transparency and preparation. Nordic Studies in Mathematics Education, 13(1), pp.49–66.

Eshet-Alkalai, Y. (2004). Digital Literacy: A Conceptual Framework for Survival Skills in the Digital Era. Journal of Educational Multimedia and Hypermedia, 13(1), pp.93–106.

European Commission. (2021). Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (AI Act). Brussels: European Commission.

Gilster, P. (1997). Digital Literacy. New York: Wiley.

Gopnik, A., Sobel, D.M., Schulz, L.E. & Glymour, C. (2004). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. Developmental Psychology, 40(2), pp.162–176.

Greenhow, C. & Askari, E. (2021). Learning and teaching with social media: AI writing assistants and implications for critical thinking. British Journal of Educational Technology, 52(3), pp.1104–1118.

Hmelo-Silver, C.E. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). Educational Psychologist, 42(2), pp.99–107.

Holstein, K., McLaren, B.M. & Aleven, V. (2019). Designing for Complementarity: Teacher and Student Needs for Orchestration Support in AI-Enhanced Classrooms. In: Proceedings of AIED 2019, LNCS 11625. Cham: Springer, pp.157–171.

Joshi, P. et al. (2020). State of the Art in Low-Resource NLP: Trends and Future Directions. Proceedings of ACL 2020, pp.45–54.

Kasneci, E., Sessler, K., Küchemann, S. et al. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. Learning and Individual Differences, 103, 102274.

KHDA. (2023). Artificial Intelligence in Education: Guidance for Schools. Dubai: Knowledge and Human Development Authority.

Khosravi, H., et al. (2022). Explainable AI in education: A conceptual framework. Computers & Education: Artificial Intelligence, 3, 100071.

Kilpatrick, J. (2001). Adding It Up: Helping Children Learn Mathematics. Washington DC: National Academies Press.

Kirschner, P.A., Sweller, J. & Clark, R.E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41(2), pp.75–86.

Krajcik, J.S. & Blumenfeld, P.C. (2006). Project-Based Learning. In: R.K. Sawyer (ed.), The Cambridge Handbook of the Learning Sciences. Cambridge: Cambridge University Press, pp.317–334.

Krokfors, L., Kangas, M., Kopisto, K. et al. (2021). Phenomenon-based learning and assessment in Finland: A national curriculum reform evaluation. Journal of Curriculum Studies, 53(5), pp.665–685.

Lehrer, R. & Schauble, L. (2006). Cultivating model-based reasoning in science education. Cambridge Journal of Education, 36(4), pp.443–453.

Lonka, K. (2018). Phenomenal Learning from Finland. Helsinki: Edita.

Luccioni, A.S., Viguier, S. & Ligozat, A.-L. (2022). Estimating the Carbon Footprint of Machine Learning Training and Inference. arXiv:2204.05149.

Luckin, R., Holmes, W., Griffiths, M. & Forcier, L.B. (2022). Artificial Intelligence and Education: Promise and Implications for Teaching and Learning. Paris: UNESCO.

Madaan, A., et al. (2023). Testing the Limits: On Systematic Generalization in Large Language Models. arXiv:2304.15004.

Marcus, G. & Davis, E. (2019). Rebooting AI: Building Artificial Intelligence We Can Trust. New York: Pantheon.

Mhlambi, S. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. Philosophy & Technology, 33, pp.659–684.

MIT CSAIL. (2024). Benchmarking Generative AI in STEM Education. Cambridge, MA: Massachusetts Institute of Technology.

Moss, G., Jewitt, C., Levačić, R., Armstrong, V. & Cardini, A. (2007). The Interactive Whiteboards, Pedagogy and Pupil Performance Evaluation. Becta Research Report. Coventry: Becta.

OECD. (2019). OECD Learning Compass 2030: A Series of Concept Notes. Paris: OECD.

OECD. (2021). AI and the Future of Skills. Paris: OECD Publishing.

Papert, S. (1980). Mindstorms: Children, Computers, and Powerful Ideas. New York: Basic Books.

Pellegrino, J.W., Chudowsky, N. & Glaser, R. (2001). Knowing What Students Know: The Science and Design of Educational Assessment. Washington DC: National Academies Press.

Pellegrino, J.W. & Hilton, M.L. (2012). Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century. Washington DC: National Academies Press.

Perelman, L. (2014). When “the state of the art” is counting words. Journal of Writing Assessment, 7(1).

Perkins, D.N. & Salomon, G. (1992). Transfer of Learning. International Encyclopedia of Education, 2nd ed. Oxford: Pergamon Press.

Risko, E.F. & Gilbert, S.J. (2016). Cognitive offloading. Trends in Cognitive Sciences, 20(9), pp.676–688.

Scardamalia, M. & Bereiter, C. (2006). Knowledge Building: Theory, Pedagogy, and Technology. In: R.K. Sawyer (ed.), The Cambridge Handbook of the Learning Sciences. Cambridge: Cambridge University Press, pp.97–118.

Schaeffer, R., et al. (2023). Are Emergent Abilities of Large Language Models Illusory?. arXiv:2304.15004.

Seixas, P. & Morton, T. (2013). The Big Six Historical Thinking Concepts. Toronto: Nelson.

Selwyn, N. (2016). Education and Technology: Key Issues and Debates. 2nd ed. London: Bloomsbury.

Selwyn, N. (2022). Education and Technology: Key Issues and Debates. 3rd ed. London: Bloomsbury.

Siegler, R.S. (1996). Emerging Minds: The Process of Change in Children's Thinking. Oxford: Oxford University Press.

Singapore Ministry of Education. (2018). Transforming Science Practical Assessment: Policy Brief. Singapore: MOE.

Singapore Ministry of Education. (2022). Inquiry-Based Pedagogies and AI Alignment in Schools. Singapore: MOE.

Sparrow, B., Liu, J. & Wegner, D.M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. Science, 333(6043), pp.776–778.

Star, J.R. & Newton, K.J. (2009). The Nature and Development of Experts' Strategy Flexibility for Solving Equations. ZDM Mathematics Education, 41, pp.557–567.

Strubell, E., Ganesh, A. & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. Proceedings of ACL 2019, pp.3645–3650.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L. & Goodman, N.D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. Science, 331(6022), pp.1279–1285.

UAE Government. (2019). UAE National Artificial Intelligence Strategy 2031. Abu Dhabi: Government of the UAE.

UNESCO. (2023). Guidance for Generative AI in Education and Research. Paris: UNESCO.

Vuorikari, R., Kluzer, S. & Punie, Y. (2022). DigComp 2.2: The Digital Competence Framework for Citizens – with new examples of knowledge, skills and attitudes. Luxembourg: Publications Office of the European Union.

Vygotsky, L.S. (1978). Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard University Press.

Wineburg, S. (2001). Historical Thinking and Other Unnatural Acts. Philadelphia: Temple University Press.

Xu, K., et al. (2023). Small-Data Approaches to Causal Learning in Artificial Intelligence. Proceedings of NeurIPS 2023.

Zheng, B., Warschauer, M., Lin, C.-H. & Chang, C. (2016). Learning in One-to-One Laptop Environments: A Meta-Analysis and Research Synthesis. Review of Educational Research, 86(4), pp.1052–1084.

Zhu, S.-C. (2021). Theories and Representations in Vision, Language, and Intelligence. Annual Review of Vision Science, 7, pp.517–547.