The Open Measurement Layer: Toward Shared, Auditable Infrastructure for Evaluating Educational AI

Kumar Srivastava*
Quantum Learning Machines, Inc., Sammamish, Washington
kumar@quantumlearningmachines.com

4 August 2026

Abstract

Generative AI has severed the assumption that underwrites most educational assessment: that a strong product indicates a strong learning process. The field's most considered response is to move assessment toward richer, process-based evidence gathered over time. I argue that this response is correct but incomplete, because process-based assessment at scale is AI-mediated assessment, and the measurement of educational AI itself remains overwhelmingly closed, vendor-controlled, and difficult to reproduce. An ecosystem that has opened its content and its tools while leaving judgment proprietary has not yet opened the layer on which the credibility of the others depends. This paper defines an *open measurement layer* comprising five components: shared construct ontologies, open evaluators, contamination-controlled benchmarks, limitations-first documentation, and interoperable instrumentation. Openly licensed artifacts that instantiate each component show the layer to be buildable today. The paper closes with open problems, including governance of held-out evaluation sets, documentation standards for evaluators, and incentive design, whose resolution will require coordination among institutions rather than action by any single organization.

Keywords: educational assessment; artificial intelligence in education; measurement validity; open educational infrastructure; benchmarking; data contamination; open science

1 Introduction: the assessment crisis is a measurement problem

Most educational assessment rests on an assumption so load-bearing that it was rarely stated: a strong product (the essay, the proof, the problem set) indicates a strong process of learning. Grading the artifact was a reasonable proxy for evaluating the learner because, for most of educational history, only a learner who had done the cognitive work could produce the artifact. Generative AI severs this link. A language model can now produce polished final products on demand, in any subject, at any level of apparent sophistication, leaving the artifact mute about the process behind it.

The field's responses cluster into two families. The first is a detection arms race: tools that attempt to classify text as human- or machine-written, met by paraphrasing and obfuscation

*The author founded Quantum Learning Machines, Inc., whose openly licensed artifacts are described in Section 5. A full competing-interests statement appears at the end of the paper.

techniques that defeat them. The most comprehensive independent evaluation to date tested fourteen detection tools and concluded that they are neither accurate nor reliable, with accuracy degrading sharply under machine paraphrase and manual editing, and with error patterns that expose honest writers to false accusation [14]. Detection treats the symptom, loses ground each model generation, and imposes its costs disproportionately on students least positioned to contest an accusation.

The second response is more considered: redesign assessment itself around process rather than product. A 2026 convening of researchers, technologists, and education leaders hosted by the Stanford Accelerator for Learning and ETS articulated this direction, defining responsible assessment as grounded in learners' sociocultural contexts, generating valid, trustworthy, and context-specific inferences, and resting on the accumulation of multiple sources of evidence through socially situated performance over time [12]. Parallel work at ETS has articulated principles for responsible AI in measurement, emphasizing validity, fairness, transparency, and accountability across the AI lifecycle [5].

I take this second direction to be correct, and argue that it is incomplete in a way that undermines it. Process-based evidence at scale is not gathered by hand. It is gathered, scored, and interpreted by AI systems: tutoring systems that log interaction, scoring models that evaluate open-ended work, adaptive engines that decide what a learner sees next. The redesign of assessment therefore routes trust *through* educational AI. And the evidence that a scoring model is valid, that a tutor's learner model is calibrated, or that an adaptive policy is fair is today produced almost entirely by the vendors themselves, on private benchmarks, with failure modes unreported, in formats no third party can reproduce. The assessment crisis is, one level down, a measurement crisis.

This paper names the missing layer and specifies what it requires. It makes four contributions: (1) a framework, the *open measurement layer*, comprising five components that together make the evaluation of educational AI a shared, inspectable public good; (2) a reflexivity argument: the criteria the field is converging on for responsible assessment of learners apply with equal force to the measurement of the AI systems doing the assessing; (3) an existence proof, via openly licensed artifacts, that the five components can be built today; and (4) a set of open problems whose resolution belongs to shared governance rather than to any single actor.

2 Background: two openings, and a third that has not happened

The movement toward openness in education has proceeded in recognizable stages, each opening a different layer of the stack.

The first opening was *content*. The open educational resources (OER) movement established that curricula, textbooks, and instructional materials could be openly licensed, adapted, and redistributed, culminating in the first international normative instrument on the subject, adopted unanimously by UNESCO's General Conference in November 2019 [13]. Open content is now mainstream: complete openly licensed curricula exist across subjects and grade bands, and openly licensed materials circulate globally under the Recommendation's framework.

The second opening addressed *tools and practice*. Open-source learning platforms, openly documented knowledge structures, open practice repositories, and, most recently, open-weights language models adapted for educational use have made the machinery of instruction inspectable and modifiable in ways proprietary platforms never were. A teacher, researcher, or ministry can today assemble a substantially open instructional stack: open curriculum, open platform, open models.

The third layer is *measurement*: how the quality of educational AI is judged. Here there has

been no comparable opening. The question "does this system teach well, score fairly, and model learners accurately?" is answered, in nearly every case, by the organization selling the system. Content and practice can be fully open while evaluation remains proprietary, and when it does, every quality claim in the ecosystem ultimately rests on vendor assertion. The stack is open at its inputs and closed at the point of judgment.

3 The gap: why closed judgment breaks an open stack

Consider what closure at the measurement layer does to the layers below it.

Demos substitute for evidence. Absent shared evaluation infrastructure, procurement and adoption decisions are made on demonstrations and marketing claims. A district evaluating two tutoring systems has no common instrument for comparing them; a researcher attempting to reproduce a vendor's reported gains has no access to the benchmark, the scoring rubric, or the evaluation code. The claim "our system improves outcomes" is unfalsifiable as offered.

Private benchmarks cannot be checked. When evaluation sets are proprietary, three questions become unanswerable: whether the benchmark measures the construct it claims to; whether the reported numbers were computed as described; and whether the benchmark has leaked into training data. The third question is not hypothetical. The natural-language-processing community has documented that evaluation is compromised when models train on the test splits of the benchmarks used to evaluate them: contamination inflates measured performance, is difficult to detect from outside, and can propagate wrong scientific conclusions into the literature [10]. Educational AI inherits this problem in a sharper form, because the corpora most useful for training educational models (problem sets, solutions, rubrics, exemplar essays) are precisely the materials from which benchmarks are built. A benchmark a model may have trained on is not a measurement; it is a memory test.

Evaluation does not reproduce. Traditional evidence formats quietly assume the thing evaluated holds still: the same instrument, given the same input, yields the same result within known error, which is what makes a published validation meaningful. Generative components break this assumption. The same student utterance can draw different responses from the same system on the same day, so a demonstration, a pilot, or even a validation study describes a single draw from a distribution, not the system a district deploys. Evidence about such systems must therefore characterize the distribution (variance across runs, behavior across versions) rather than report a point estimate from one sampled trajectory, and the record must preserve what the system actually did in each consequential interaction.

The reflexivity failure. The responsible-assessment literature converges on a set of criteria for judging learners: inferences should be valid, fair, transparent, and trustworthy, and should rest on multiple sources of evidence [12, 5]. The central claim of this paper is that these criteria apply reflexively, binding the measurement of the AI systems doing the assessing with the same force that they bind the assessment of learners. An evaluation of a scoring model that no independent party can inspect is not transparent, whatever its internal rigor; the same holds down the list, since a benchmark with unknown contamination status cannot support valid inference, and a learner model whose calibration has never been checked from outside has no claim to trustworthiness beyond

its vendor's word. A field that accepts these criteria for students cannot coherently exempt its instruments.

4 Five components of an open measurement layer

The layer comprises five components. For each, this section gives the definition, the reason the component matters, what "open" concretely requires, and the failure mode it prevents.

4.1 Shared construct ontologies

A construct ontology is a public, research-grounded vocabulary of what is being measured: skills, misconceptions, and prerequisite relations, each with a stable identifier. Measurement begins with agreement on constructs, and the learning sciences supply the theoretical basis for that agreement: knowledge components as acquired units of cognitive function inferable from performance on related tasks [6], and misconceptions understood not as mere errors but as systematic, productive prior conceptions with documented structure [11]. The intelligent-tutoring tradition demonstrated decades ago that inference over such component structures can be made computational [2]. For this component, openness requires open licensing of the ontology itself, public identifiers, documented provenance to the research literature, and open interfaces for programmatic access. The failure it prevents is incomparability: the situation in which every vendor measures privately defined constructs, so that cross-tool comparison, meta-analysis, and cumulative science become impossible.

4.2 Open evaluators

The evaluators are the models that judge learner or tutor output (scoring classifiers, misconception detectors, rubric appliers), together with any generative component whose behavior shapes the evidence used to judge a learner. In AI-mediated assessment the evaluator is the instrument, and a generative tutor that produces the interaction a learner is then assessed on belongs to the measurement chain whether or not it is called an evaluator. The concern is not hypothetical: systematic study of language models used as judges has documented position bias, verbosity bias, and self-enhancement bias, among other failure modes, in exactly the configuration (a model scoring open-ended output) on which AI-mediated assessment relies [15]. An instrument that cannot be inspected cannot be checked for construct validity, bias, or failure modes, and its judgments cannot be contested by those they affect. Openness here means open weights, training-data documentation sufficient to assess representativeness and leakage, and versioning that ties every score to the exact evaluator that produced it; where weights cannot be opened, it means runtime behavior fully recorded (Section 5). The failure prevented is unauditable judgment: consequential decisions about learners flowing from models no independent party can examine.

4.3 Contamination-controlled benchmarks

Contamination-controlled benchmarks are evaluation sets engineered for integrity: held-out splits with governed access, versioning, and design features that make later contamination detectable. The need follows from Section 3: a compromised benchmark measures memory rather than capability, and contamination is difficult to detect from outside [10]. Openness for this component is not the same as full publication; it is open *governance*: public documentation of the benchmark's construction and construct mapping, in the spirit of recent efforts to make large-model evaluation transparent and multi-dimensional [7], together with public versioning, disclosed contamination-control methodology,

and access rules administered by a neutral steward rather than by any evaluated party. For generative components, one further requirement follows from Section 3: performance must be reported as a distribution across runs, with model versions and sampling parameters pinned, because a single-run point score on a non-deterministic system is a sample, not a measurement. The failure prevented is inflated self-reported scores that no one can independently deflate.

4.4 Limitations-first documentation

Limitations-first documentation is standardized documentation for models and datasets that foregrounds failure modes, disaggregated performance, and out-of-scope uses, with numbers traceable to the evaluations that produced them. The templates exist: model cards for trained models [8] and datasheets for datasets [3]. The educational stakes are also documented: algorithmic bias in educational systems has been observed across race, gender, disability status, and language background, and aggregate accuracy conceals exactly these disparities [1]. Adoption of the templates must therefore come with a stricter norm than current practice: the limitations section carries the document's central information rather than serving as a disclaimer, and unflattering numbers are reported with the same precision as flattering ones. The failure prevented is a capability claim that stays silent about the learners for whom the system fails.

4.5 Interoperable instrumentation

Interoperable instrumentation means shared event schemas and measurement interfaces that let evaluation travel across platforms, so that evidence generated in one system can be interpreted, audited, and aggregated by another. The precedent exists: the Experience API, developed in the distributed-learning community and standardized as IEEE 9274.1.1-2023, demonstrates that cross-platform learner-experience data interchange can be specified and adopted at scale [4]. Without interoperability, every measurement result is stranded inside the platform that produced it, and the layer fragments back into proprietary silos. Openness requires open specification of event and evidence schemas, conformance tests, and crosswalks to existing standards rather than replacements for them. The failure prevented is evaluation as product: capture of the measurement layer by whichever platform hosts it.

The five components are complementary, and the layer needs all of them. An open ontology does little if the models that score against it stay closed. Open evaluators can still post inflated numbers if nothing controls contamination of the sets they are scored on. Even trustworthy scores mislead when their failure modes go undocumented, and everything above remains stranded in single platforms unless the instrumentation interoperates.

5 An existence proof

A framework paper owes the reader evidence that its proposal is buildable rather than aspirational. This section describes openly available artifacts, released by Quantum Learning Machines, that instantiate the five components. The description is deliberately architectural (interfaces, licenses, and verifiable properties) and public figures only. The section demonstrates feasibility and openness; it makes no efficacy claims (see the scoping paragraph below).

Ontology. An openly licensed (CC BY 4.0) misconception ontology spanning nine K–12 domains (physics, biology, chemistry, mathematics, computer science, earth science, English language arts,

health, and history), comprising 423 constructs. Each entry documents the student belief, the scientific reality, a counterexample, a diagnostic question, frequency, persistence and severity ratings, provenance to the research literature, and a teacher intervention. The ontology is served through a zero-authentication open API and ships with crosswalks to an open practice service (118 mappings from mathematics misconceptions to practice assignments) and an open knowledge graph (31 mappings to standards codes), so the identifiers resolve against infrastructure the community already uses. A companion learning graph of 1,884 directed edges encodes prerequisite relationships, shared concepts, and misconception overlap among the platform's simulation modes [9].

Contamination control in practice. Each ontology entry carries an explicit tier. Public-tier entries release every field; evaluation-tier entries publish the construct, the student belief, the scientific reality, and the intervention, while withholding their trigger patterns and diagnostic questions from the export so that those items remain usable as uncontaminated evaluation material. This is the requirement of Section 4.3 operating in a shipped artifact rather than a proposal: the ontology is open, and the evaluation-critical fields are governed. Openness and benchmark integrity are reconciled at the level of the individual field rather than the dataset.

Evaluators. Openly licensed evaluation classifiers for a subset of the ontology's constructs, released with limitations-first cards in the model-card format [8]: disaggregated performance, known failure modes, and out-of-scope uses are reported with the same precision as headline metrics, including figures that are unflattering. The accompanying dataset documentation follows the datasheet template [3].

Instrumentation and auditability. An openly licensed (Apache 2.0) software development kit, `qlm-measure`, specifying an evidence-event schema for measurement interactions, designed for crosswalk compatibility with IEEE 9274.1.1-2023 rather than as a replacement for it. The schema treats the instrument as a first-class subject of the record: alongside learner-side events, it carries instrument-side provenance (model identity, sampling temperature, prompt hashes, fallback occurrences, and validation status), so that when a generative component participates in an interaction, what that component did is itself part of the auditable trail. The SDK includes a *record verifier*: a tool that checks the structural integrity of a measurement record, meaning its tamper-evidence and the internal consistency of its posterior chain, without requiring access to the estimation model that produced the record. The property this establishes is auditability decoupled from trust: a third party can verify that a measurement trail is complete, unaltered, and internally consistent even though the estimator itself is proprietary. I regard this decoupling as a pragmatic resolution of an apparent conflict between open measurement and commercial model development: the *judgment trail* can be fully open even where the *estimator* is not, and the layer's integrity requirements attach to the former.

Scoping. Two limits of this section should be stated plainly. First, it is an existence proof of feasibility and openness: the artifacts show that the five components can be built and released openly today, not that any particular tool improves learning outcomes; efficacy claims require controlled studies that are outside this paper's scope. Second, the estimation methods behind the organization's systems are deliberately not described here; the paper's argument concerns the openness of the measurement layer's *interfaces, benchmarks, documentation, and audit trails*, which, as the record verifier illustrates, does not depend on disclosure of any party's estimators.

6 Relationship to responsible-assessment frameworks

The convenings and research programs now defining responsible assessment articulate principles: assessment should yield valid, trustworthy, context-specific inferences from multiple sources of evidence [12], and AI used in measurement should be developed under explicit commitments to fairness, transparency, accountability, and privacy across its lifecycle [5]. This paper is positioned one level below those documents: it specifies the shared technical infrastructure that would make the principles checkable rather than asserted. Inspectable evaluators give transparency operational content; contamination-controlled, openly governed benchmarks do the same for validity. Fairness becomes checkable when disaggregated performance is a first-class documented artifact, and trustworthiness when measurement records can be verified by parties other than their producer. The principles literature says what responsible measurement must achieve; the open measurement layer is a concrete proposal for how an ecosystem, rather than a vendor, achieves it.

7 Open problems and an invitation

Five problems are unresolved, and their common property is that no single organization can or should resolve them alone.

Contamination control at ecosystem scale. Held-out evaluation data must simultaneously stay out of training corpora and remain available for legitimate evaluation, a tension that sharpens as training pipelines ingest ever more of the public web. Detection methods, canary conventions, and disclosure norms all need development beyond the current state of the art [10].

Governance of shared benchmarks. Someone must hold the held-out sets. The holder acquires power over every evaluated system, so the steward must be neutral, plural, and accountable—a standards body, a consortium with rotating custody, or a purpose-built institution. The governance design is as important as the benchmark design.

Standards for evaluator documentation. Model cards and datasheets are necessary but not sufficient for evaluators specifically: an instrument that judges learners needs documentation norms about construct mapping, error consequences, and contestability that the general-purpose templates do not yet specify.

Interoperability convergence. The layer should extend existing standards, not proliferate new ones; the path from the current specification landscape to a shared evidence schema requires deliberate crosswalk work with the standards community.

Incentive design. Vendors currently benefit from closed evaluation. Adoption of open measurement will follow procurement requirements, funding conditions, and publication norms that reward verifiable claims over asserted ones—levers held by districts, ministries, funders, and journals, not by any technology builder.

Some measurement layer will be built regardless of what the field decides, because every organization shipping educational AI also ships, implicitly, a way of judging it. Left to that default, evaluation accretes as a set of proprietary regimes, and the closure that content and tools escaped over the past two decades is reproduced at the layer that determines what counts as quality. This paper has argued

for the alternative: measurement as shared infrastructure, jointly governed and openly inspectable, specified in five components that Section 5 shows to be buildable now. The problems that remain are institutional rather than technical, and they sit with the districts, ministries, funders, standards bodies, and journals whose choices determine which evidence counts. The artifacts described here are released under open licenses as a starting contribution to that work.

Data and code availability

The ontology described in Section 5 is available under CC BY 4.0 at https://play.quantumlearningmachines.com/api/labpath/dataset-export; the counts reported in this paper correspond to release v1.0. The `qlm-measure` SDK is available under Apache 2.0; the instrument-side provenance fields described in Section 5 are present as of schema v0.2.0.

Competing interests

I founded Quantum Learning Machines, Inc., which produced the artifacts described in Section 5, and I hold a financial interest in its success. The framework advanced in this paper would, if adopted, apply to that organization's systems on the same terms as to any other party's. One structural interest deserves explicit note: the specification in Sections 4 and 5 permits estimation models to remain proprietary provided the judgment trail around them is open, and Quantum Learning Machines benefits from that permission. The argument for that decoupling is given in Section 5, and readers should weigh it with this interest in view.

References

[1] R. S. Baker and A. Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 32:1052–1092, 2022. https://doi.org/10.1007/s40593-021-00285-9

[2] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4:253–278, 1995.

[3] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. https://doi.org/10.1145/3458723

[4] IEEE. *IEEE Standard for Learning Technology—JavaScript Object Notation (JSON) Data Model Format and Representational State Transfer (RESTful) Web Service for Learner Experience Data Tracking and Access*. IEEE Std 9274.1.1-2023, IEEE Standards Association, 2023.

[5] M. S. Johnson et al. *Responsible AI for Measurement and Learning*. ETS Research Report RR-25-03, Educational Testing Service, 2025.

[6] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science*, 36(5):757–798, 2012. https://doi.org/10.1111/j.1551-6709.2012.01245.x

[7] P. Liang, R. Bommasani, T. Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. https://doi.org/10.48550/arXiv.2211.09110

[8] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*, pages 220–229. ACM, 2019. https://doi.org/10.1145/3287560.3287596

[9] Quantum Learning Machines. *QLM LabPath Dataset v1.0* [Data set], 2026. License: CC BY 4.0. https://play.quantumlearningmachines.com/api/labpath/dataset-export

[10] O. Sainz, J. A. Campos, I. García-Ferrero, J. Etxaniz, O. Lopez de Lacalle, and E. Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787. ACL, 2023. https://doi.org/10.18653/v1/2023.findings-emnlp.722

[11] J. P. Smith III, A. A. diSessa, and J. Roschelle. Misconceptions reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, 3(2):115–163, 1994. https://doi.org/10.1207/s15327809jls0302_1

[12] Stanford Accelerator for Learning and ETS. *Responsible Assessment in the AI Era: Key Insights from a Future-Focused Convening*. White paper, Stanford University, 2026. https://acceleratelearning.stanford.edu/app/uploads/2026/07/ResponsibleAssessmentintheAIEra_StanfordAcceleratorforLearning.pdf

[13] UNESCO. *Recommendation on Open Educational Resources (OER)*. Adopted by the General Conference at its 40th session, 25 November 2019 (Document 40 C/32).

[14] D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltýnek, J. Guerrero-Dib, O. Popoola, P. Šigut, and L. Waddington. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19:26, 2023. https://doi.org/10.1007/s40979-023-00146-z

[15] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. https://doi.org/10.48550/arXiv.2306.05685