OPEN MEASUREMENT

The Open Measurement Layer

Why open education AI needs one more layer — and what it's made of

Kumar Srivastava · Quantum Learning Machines · v1.0 · July 2026

Education AI is having its open moment. Open curricula (OpenSciEd, Illustrative Mathematics) tell tools what to teach. Open knowledge graphs give them shared structure. Open services and open-weights models give teachers real alternatives to black boxes. A serious open stack for teaching is assembling in public, and the biggest players are now building on it.

One layer is still missing, and it is the layer that decides whether the rest can be trusted: open measurement.

Today, the question "does this AI tool actually teach well?" is answered by the vendor selling it. Demos substitute for evidence. Benchmarks, where they exist, are private, contaminated, or self-graded. A district choosing between AI tutors has open curricula to point them at content — and marketing to tell them about quality. Open inputs, closed judgment.

In early 2026, the Stanford Accelerator for Learning and ETS convened the field around exactly this problem and published Responsible Assessment in the AI Era, arguing that assessment must move beyond end-of-process testing toward richer evidence of learning — and that it must remain valid, fair, transparent, and trustworthy. We agree, and we want to add the corollary that follows from building this infrastructure rather than only theorizing it: those same four standards have to apply to how we measure the AI, not only to how the AI measures students. An assessment ecosystem that is open everywhere except at the point of judgment is open in inputs and closed where it counts.

The fix is the same move the field already made twice. Open curricula opened what is taught. Open platforms opened where practice happens. The third layer opens how quality is known. We call it the open measurement layer, and it has five components:

1. Shared ontologies of learning constructs. Public, research-cited vocabularies of what there is to measure — misconceptions, skills, prerequisite structure — so that tools, researchers, and benchmarks can talk about the same things. A misconception detector is only comparable to another if both are pointed at a common map.

2. Open evaluators. The models that judge AI output — misconception detectors, pedagogy scorers, restraint classifiers — released as open weights with documented training data, so anyone can run the judge, inspect the judge, and disagree with the judge.

3. Contamination-controlled benchmarks. Evaluation sets built for integrity from day one: held-out splits that never ship, evaluation tiers with detection signals redacted, canary strings, versioned releases. A benchmark a model may have trained on is a press release, not a measurement.

4. Limitations-first documentation. Model cards and dataset cards that lead with what the artifact cannot do, with every number traceable to a versioned result. The test of a card is whether a skeptic can check it.

5. Interoperability and open instrumentation. Measurement artifacts published in formats that plug into the open knowledge graphs, curricula, and practice platforms already in classrooms — and open SDKs that let any tool emit, run, and consume measurement — so evaluation becomes plumbing, not product.

This layer is not hypothetical. Our pieces of it are live today: a 423-construct, research-cited misconception ontology (CC-BY-4.0, open API, with learning-graph and standards-alignment layers); two released evaluation classifiers with limitations-first cards — including honest, unflattering numbers, published because a measurement company that hides its own measurements is a contradiction; an openly licensed tutoring model documented the same way; and an evaluation-tier design that lets one public dataset serve honestly as both training data and benchmark. Interop with the open knowledge-graph ecosystem is in progress. Others hold other pieces. The point of naming the layer is that no one company should own it — it should be built the way open curricula were built: in public, by many hands, held to common standards.

The lineage matters. This is not a new idea; it is the OER movement's third act. Open content came first. Open tools and practice followed. Open measurement completes the loop — because a stack that is open everywhere except at the point of judgment is open in inputs and closed where it counts.

Where our open line sits — in daylight. We are direct about what is open and what is not. Open, and yours to keep: the misconception ontology (CC-BY), the evaluation models and their limitations-first cards, the evidence-record format, and the SDK verifier that audits a measurement trail without our involvement. Not open-source, and operated by us: the estimation engine — the calibrated Bayesian and

psychometric machinery that turns evidence into proficiency estimates. That engine is how we sustain the company, and we would rather say so plainly than blur the line. What we can do is make its workbench free to the people who most need rigorous measurement: researchers. We built this measurement discipline to hold ourselves accountable first — it is why our own model cards lead with unflattering numbers — and we are making that same instrument free for researchers to measure their work rigorously and productively, whatever tool they are studying. Open where the field needs shared infrastructure; free where researchers need capability; honest about the boundary between them.

An invitation. If you build teacher tools: publish your evaluations, not just your features. If you fund education AI: require the five components above before believing a quality claim. If you research learning: the ontologies and evaluators are yours to use, extend, and refute — telling us where they are wrong is a contribution. And if you are a district: there is a five-question checklist version of this document, written for procurement. Ask every vendor all five.

When the measurement layer is open, “which tool is good” stops being a marketing question. That is the ecosystem students and teachers deserve.

Cite as: Srivastava, K.S. (2026). The Open Measurement Layer, v1.0. Quantum Learning Machines.
quantumlearningmachines.com/resources · Data & models: play.quantumlearningmachines.com/developer · huggingface.co/QuantumLearningMachines