Axiom Flow

TECHNICAL WHITE PAPER

Measuring Understanding When AI Can Produce the Answer

A technical white paper on Axiom Flow

Rajika Chathuranga

axiomflow.app

September 2026

Summary

Generative AI has broken the link between what a student submits and what a student knows. Assessment that grades an artifact can no longer tell the difference between a student who understands the material and one who obtained a good artifact.

Axiom Flow addresses this by changing what the task is rather than by policing how it is done. Students teach an AI student called Sam, who begins each session holding deliberately constructed misconceptions about the source material. Sam then sits a multiple choice exam using only what the student taught him. Sam's score becomes the student's result.

Three properties make this different from an AI tutor or a quiz generator:

The grade is not produced by an AI. Sam's answers are compared to the correct answers by direct string match. No language model judges the student's work.

The unit of assessment is a transferred concept. Each misconception maps to exactly one exam question. A wrong answer identifies precisely which idea failed to transfer.

The process is fully recorded. Sam's complete mental state is stored after every teaching turn, so an instructor can see not just the final result but the path a student's explanation took.

Axiom Flow has run as a graded component across fourteen assignments in a second year Computer Systems Programming module at the Department of Computer Engineering, University of Peradeniya, with 100 students.

This paper describes how the system works, what a survey of that cohort shows, and, importantly, what we do not yet claim. Axiom Flow has not been validated against conventional assessment outcomes. That study is described at the end of this paper as the necessary next step.

1. The problem is measurement, not cheating

The instinctive institutional response to generative AI has been an integrity response: detect it, restrict it, penalise it. Detection tools are unreliable, restriction is unenforceable outside supervised conditions, and both consume effort that does not improve assessment.

Assessment scholarship has largely moved past this framing. The argument, advanced by Dawson, Corbin, Liu, Bearman and others, is that AI is a validity problem before it is an integrity problem. When a student can submit work that appears to demonstrate competence without possessing it, the assessment has failed at its stated job of measuring learning. The useful question is not “did they cheat” but “does this task still measure what it claims to measure.”

The same literature distinguishes *discursive* responses, which add rules and declarations students can ignore, from *structural* responses, which change the mechanics of the task itself. Discursive changes produce what has been called an enforcement illusion: a policy label attached to a task that is otherwise unchanged.

The evidence for urgency

The clearest empirical picture comes from Strömberg, Lei and Wu (2026), who tracked 26,811 Chinese secondary students in grades 7 to 12 across nine subjects over 30 months, exploiting staggered AI adoption in a difference in differences design.

After adoption:

- Homework scores rose 18 percent while completion time fell 30 percent
- Monthly closed book exam scores fell 20 percent within six months
- Entrance exam scores fell 18 and 24 percent, with the full penalty only emerging after roughly two years

The detail that matters most for assessment design is the distribution of those losses. They were concentrated among roughly 80 percent of AI users whose behaviour matched homework outsourcing, identified by unusually short completion times paired with high homework scores. Students who used AI but kept completion times similar to non users showed only small losses.

The problem, in other words, is not AI use. It is outsourcing. And conventional assignments cannot distinguish the two, because a submitted artifact looks the same either way. The gap also took two years to become fully visible, which means an institution relying on coursework marks would see improvement long before it saw the damage.

Axiom Flow's design follows directly from this. The goal is not to prevent students from using AI. It is to build a task where using AI to understand the material helps, and using AI to produce the answer does not.

---

2. What Axiom Flow is

Axiom Flow is an assessment layer. It is not a tutor, a quiz generator, a content platform or a learning management system. It integrates with an existing LMS through LTI, or runs standalone through its own portal for instructors without one.

A session works as follows.

Setup. An instructor or student provides source material: a PDF, pasted text, or a YouTube URL. This defines the knowledge boundary for the session. Nothing outside it is assessed.

Misconception design. An authoring agent, Atlas, analyses the material and constructs a set of misconceptions, configurable at 5, 10, 15 or 20. Each is a specific, plausible, wrong belief about a concept in the material. Atlas then generates exactly one exam question per misconception.

Teaching. Sam begins the session holding those misconceptions as his own beliefs. He has no independent way to verify what is true. The student works through them one at a time, explaining the concept in their own words. Sam asks clarification questions when an explanation is incomplete, and updates the belief in question when it is corrected.

Exam. Sam sits the multiple choice exam using only the beliefs he now holds. Any misconception the student failed to correct produces a wrong answer.

Result. The student receives a score, plus a gap report identifying which specific ideas did not transfer and why.

The position on AI use

Students are free to use ChatGPT or any other tool to understand the material. Nothing is blocked, detected, or penalised.

What is blocked is pasting into the teaching interface. The explanation given to Sam has to be composed by the student, typed or spoken.

This is a deliberate and limited claim. It does not make outsourcing impossible; a student can read a generated explanation and retype it, and Section 6 documents a student in our own cohort doing exactly that. What it does is remove the frictionless path, and, more importantly, produce a transcript. When a student transcribes an explanation they do not understand, Sam's follow up questions expose it, because answering them requires understanding the explanation rather than possessing it.

---

3. Why teaching is a reasonable basis for assessment

The idea that explaining something reveals whether you understand it is old and intuitive. The research base is more specific than the intuition, and worth stating precisely, including where it is weaker than commonly assumed.

Teachable agents are an established paradigm. The direct ancestor of this design is Betty's Brain, developed at Vanderbilt and Stanford, in which students teach an agent named Betty by building a causal model of a science topic, while a second agent, Mr. Davis, generates quizzes and mentors the student. That two agent division of labour, one agent to be taught and one to author and assess, maps onto Sam and Atlas. It has roughly two decades of published study behind it.

Students work harder for an agent than for themselves. Chase, Chin, Oppezzo and Schwartz (2009) named this the protégé effect. Holding the software constant and varying only whether students believed they were teaching an agent or learning for themselves, the teaching group spent more time on learning activities and learned more. The effect was most pronounced among lower achieving students.

The effect sizes are real but modest. Kobayashi's meta analysis of 39 studies found a weighted mean effect of $g = 0.27$ for teaching after study. The important finding was a moderator: teaching after studying with the expectation of teaching gave $g = 0.48$, while teaching without that expectation was indistinguishable from zero. Interactive teaching outperformed non interactive. Axiom Flow satisfies both conditions, since the student knows in advance they will teach Sam and the teaching is a dialogue.

Explaining alone is not enough, and this shapes the design. Roscoe and Chi's review found that tutors exhibit a pervasive knowledge telling bias: even when trained, they summarise source material rather than elaborate on it, and consequently the potential benefit of tutoring

is rarely realised. Their empirical work found that the behaviour that does produce understanding, reflective knowledge building, was most often triggered by the tutee, specifically when the tutee asked a question requiring an inference.

This is why Sam asks questions. It is not a conversational nicety. It is the mechanism that converts summarising into understanding, and it is supported by direct experimental work on synthetic tutees by Shahriar and Matsuda, who found that follow up questions from an artificial tutee facilitated tutor learning and were particularly effective for students with low prior knowledge.

Recent LLM deployments confirm both the promise and the risk. Rogers and colleagues (CHI 2025) deployed an LLM teachable agent across four experiments with 119 university students, and found it conveyed a novice persona comparably to a human and delivered comparable benefits, including more accurate student self assessment. Wang and colleagues (L@S 2026) deployed a teachable agent to 546 students in an undergraduate algorithms course over eleven weeks, analysing 3,809 dialogues. They found explanation oriented behaviours associated with conceptual understanding, and they also observed direct reuse of externally sourced content. That published observation is the documented reason Axiom Flow blocks pasting.

---

4. How the system works

Sam’s mind is a data structure, not a conversation

The central design decision is that Sam’s understanding is not implicit in a chat history. It is an explicit, stored list of belief nodes, one per misconception:

```
Thought
  id                    T1, T2, ...
  content               the belief, in Sam's own voice
  resolved_percentage   0 to 100
```

Every node carries its full revision history, so each belief has a traceable evolution from the original misconception to whatever the student left it as.

This mirrors the causal map in Betty's Brain, where the agent's knowledge was visible as a graph the student had built. Making the agent's knowledge externalised and inspectable is what makes the result interpretable rather than a black box. When a student loses a mark, there is a specific stored belief that caused it, and a specific conversation that produced that belief.

Teaching is sequential and targeted

Students work one belief at a time. Each turn addresses a single node, and Sam may revise only that node. He is constrained during teaching by explicit rules:

- He cannot use technical terms the student has not introduced, and must ask what they mean
- If told only that he is wrong, he does not deduce what is right; he asks
- If a student's correction addresses only part of a belief, he retains the uncorrected part and continues to hold it

The exam is knowledge isolated

At exam time Sam operates under five constraints:

1. He may use only his internal beliefs, not general knowledge
2. If a belief says something is wrong without saying what is right, he cannot select the right answer
3. He may not guess by eliminating options
4. If a belief leads to a wrong answer, he must choose that wrong answer
5. If every option contradicts an explicit detail he holds, he skips the question

Section 7 states the limits of this mechanism honestly.

The grade is arithmetic

Sam's selected option is compared to the correct answer by direct string match. One point per correct answer. No language model evaluates the student, and no language model assigns the grade.

This is the most important sentence in this paper for anyone responsible for assessment policy. The role of AI in Axiom Flow ends at authoring the questions and simulating the student. The measurement itself is deterministic and reproducible: the same beliefs and the same exam yield the same score every time.

The exam is single attempt. A score cannot be improved by repetition.

Question design draws on concept inventory methodology

The distractors are not filler. For each question, the authoring agent is required to produce one distractor that exactly matches the logic or false values of the corresponding misconception, acting as a direct trap for a student who left that idea uncorrected, and one near miss distractor that is conceptually close to the truth but technically flawed, to catch partial understanding. Question stems must state the mechanism being tested rather than asking which statement is true.

This follows the approach of research grade concept inventories such as the Force Concept Inventory, where distractors are drawn from documented student misconceptions so that the pattern of wrong answers is diagnostic rather than merely incorrect. The important difference, stated plainly in Section 7, is that concept inventories are validated over years and Axiom Flow generates them in seconds.

The gap report is separate from the grade

After the exam, an auditing pass compares Sam's final beliefs against the source material and identifies the specific phrases in his thinking that are factually wrong, along with corrections. Source material is authoritative for figures and definitions. Omissions are never flagged as errors.

This audit produces the student's feedback. It has no effect on the score.

What instructors can see

Every teaching turn stores a complete snapshot of Sam's mental state and a record of which beliefs changed. The exam stores the exact state of Sam's beliefs at the moment he sat it.

Instructors therefore have access to the full teaching transcript, how each misconception was attempted, which ones survived, the evolution of each belief over time, and session viewing and active engagement time.

This is process evidence of a kind conventional assignments cannot produce. A submitted essay shows an endpoint. A teaching transcript shows how a student's explanation of an idea changed over twenty minutes of being questioned about it.

5. Deployment

LTI, for institutions. Axiom Flow runs as an LTI integration, currently live with Moodle. Sessions link to assignments and student progress records, with enforcement for deadlines, document purging, seat provisioning and institutional access rules. Blackboard, Canvas and Google Classroom are planned.

Standalone, for individual instructors. Instructors without LMS access create assignments in the Axiom Flow portal and view results there.

Self directed, for students. Students can run the same flow on their own material for revision. Results sync nowhere.

Session sharing. A session can be shared by link, creating individual copies for each participant, with the owner seeing each participant's latest score and activity.

Three model providers are used with automatic failover so that a provider outage does not interrupt an assessment. Usage is metered per operation with full token and cost accounting, billed to the institution for institutional sessions.

6. Evidence from deployment

Axiom Flow has been used as a graded component across fourteen assignments in a second year Computer Systems Programming module in the Department of Computer Engineering, University of Peradeniya, Sri Lanka, with 100 students and one instructor.

A voluntary survey drew 52 responses, a 52 percent response rate. Voluntary surveys over-represent engaged students, so the figures below should be read as favourable rather than representative.

Teaching Sam helped more than studying alone	45 yes, 7 unsure, 0 no
Compared to a regular quiz or assignment	23 much better, 23 better, 6 the same, 0 worse
Sam's questions during teaching	45 about right, 6 too many, 1 too few
The score felt like a fair reflection of understanding	**32 yes, 16 unsure, 4 no**
Clarity of what to do (1 to 5)	mean 3.8; 22 of 52 rated 3 or below

What students valued

The free responses describe the intended mechanism in students' own terms. One wrote that Sam asked logical questions when given partially correct answers, which sent them to find accurate descriptions and return to teach him, learning the material properly in the process. That is knowledge building elicited by tutee questioning, which is precisely the behaviour Roscoe and Chi identified as the difference between tutoring that works and tutoring that does not.

Another wrote that Sam pointed out things they had not known they did not know. Improved self assessment accuracy was also the primary reported benefit in the Rogers deployment.

Several described enjoying watching Sam sit the exam on the strength of their teaching, and one cited the inability to paste text as a positive feature.

Where it fell short

Perceived fairness is the weakest result, and the most important. Just under 62 percent felt the score fairly reflected their understanding. Fifteen percent were unsure or disagreed. For a measure intended for a gradebook, that number needs to be higher, and the free text identifies three specific causes.

The first is a misleading progress indicator. One student described a belief showing 97 percent resolved, proceeding on that basis, and scoring 9. The internal progress figure is set by Sam himself and can also be set directly by the student, which means it indicates navigation through the session, not readiness. It should never have been presented in a way students could read as a predictor of their score. It is not used in grading, and the interface is being changed.

The second is attribution ambiguity. Students could not always tell whether a lost mark reflected their failure to explain or Sam's failure to interpret. This is inherent to the design and cannot be eliminated, but the gap report can do more to show which belief produced which answer.

The third is question opacity. During teaching, Sam is steered toward the concept that will be assessed. Some students perceived this as him asking for something specific without saying what.

Onboarding is not good enough. Fewer than 60 percent rated task clarity at 4 or above. The concept of teaching a deliberately wrong AI student is unfamiliar and the current interface does not explain it well enough.

Cross session memory. Students objected that Sam does not retain terms taught in earlier assignments, forcing repeated explanation. This is knowledge isolation working exactly as designed, experienced as a defect. The isolation that makes each session a clean measurement is the same property that makes the fourteenth session feel repetitive. This is a genuine tension in the design, not a bug, and we have not resolved it.

The finding we think matters most

One student wrote:

> *"Sometimes I have to complete this task without a proper understanding of the lesson. Because of the workload, but the deadline is approaching. As a result, I copy the question into an AI tool and type the answer it gives without fully understanding what it contains. I know this isn't good, but when the deadline is close, I feel like I have no other option."*

This is a student in this cohort describing the outsourcing behaviour the Strömberg study identified, occurring inside Axiom Flow, despite pasting being blocked.

We include it because any institution evaluating this system should know it. Blocking paste raises the cost of outsourcing and forces the student to process the text at least once. It does not make outsourcing impossible, and we do not claim it does.

Two observations follow. First, this student was under workload pressure, which points at deployment design as much as at the tool: fourteen graded assignments in one module creates the exact conditions the confession describes. Second, and more usefully, transcription is visible in the record in a way it is not in a submitted document. When a student transcribes an explanation they have not understood, Sam's follow up questions expose it, because answering a question about an explanation requires understanding it. That signal exists in the transcript for an instructor to read.

---

7. What we do not claim

Institutions evaluating assessment technology are entitled to a clear statement of limits.

Axiom Flow has not been validated against conventional assessment outcomes. This is the most important limitation in this paper. All of the research cited in Section 3 treats learning by teaching as an *intervention that improves learning*. Axiom Flow uses it as a *measurement instrument*, where an agent's score becomes a student's grade. No published study validates

that step. The survey data in Section 6 measures acceptance and perceived value, not learning. Perceived learning and actual learning are known to diverge. Section 8 describes the study required to close this gap.

Evidence comes from a single context. One module, one instructor, one institution, one discipline. Systems programming is also a demanding case for this design, because a significant portion of the knowledge is procedural and syntactic rather than conceptual, and misconception based assessment fits conceptual knowledge best. Performance in other disciplines is unknown.

Question quality is not calibrated. Research grade concept inventories take years of empirical work to validate distractors against documented student errors. Axiom Flow generates them in seconds from arbitrary material. We do not yet have evidence on whether difficulty is stable when the same material is processed twice, and instructors should review generated questions rather than assume them.

Knowledge isolation is enforced by instruction, not by architecture. The underlying models retain their pretrained knowledge and are instructed not to use it. Alternative approaches remove the knowledge outright through machine unlearning. Prompt level isolation is a weaker guarantee. It is measurable, and we intend to measure it, but at present it is a design intention rather than a proven property.

Model variation affects comparability. Automatic failover between providers protects availability, but two students working on identical material could in principle be assessed under different underlying models. Agreement between providers is not yet quantified.

Progress indicators are not measures. The internal resolution percentage indicates position in a session and can be set by the student. It is not used in grading and should never appear in a gradebook.

Outsourcing is reduced, not eliminated. See Section 6.

---

8. Validation plan

The stored data makes the necessary studies straightforward, because Axiom Flow retains a complete snapshot of Sam’s beliefs after every teaching turn and at the moment of the exam. Any past session can be replayed exactly.

Criterion validity. The essential study. Take a cohort using Axiom Flow on material also covered by a conventional exam, and correlate the two scores per student. A meaningful positive correlation is the evidence institutions need in order to trust an Axiom Flow score in a gradebook. This requires instructor cooperation to link results and is the priority for the next academic term.

Leakage measurement. Replay stored initial belief states, where every misconception is intact and nothing has been taught, through the exam, and count how often Sam still answers correctly. That number quantifies how much pretrained knowledge escapes the isolation instructions. It can be computed today from existing records with no new student sessions.

Cross provider agreement. Replay stored exam time belief states through each model provider and compare scores. Also computable from existing records.

Question stability. Regenerate misconceptions and questions from the same material repeatedly and measure variation in difficulty.

Trajectory analysis. Turn level snapshots make it possible to reconstruct how each student's explanation moved each belief over time, and to test whether particular teaching behaviours predict transfer. This is the process evidence assessment researchers ask for and that few systems are able to produce.

We will publish these results whether or not they are favourable.

---

9. Conclusion

Generative AI has made the artifact an unreliable proxy for understanding. Detection does not fix this, and rules that students can ignore do not fix it either. What changes the situation is a task where producing the output requires the understanding.

Axiom Flow is one attempt at such a task. A student who understands a concept can correct Sam's belief about it and answer his questions. A student who possesses a correct explanation without understanding it can transcribe it, but struggles when Sam asks what it means. The resulting score is computed arithmetically, maps to specific concepts, and is accompanied by a complete record of how the student got there.

The design rests on two decades of teachable agent research. The use of that design as a graded measurement instrument does not, and we have said so plainly. Deployment across fourteen assignments with 100 students shows strong student acceptance and identifies real problems, including one student's candid account of outsourcing under deadline pressure.

The correlation study in Section 8 is what would turn this from a well grounded design into a validated instrument. We are seeking institutional partners to run it.

References

Bearman, M., Tai, J., Dawson, P., Boud, D., & Ajjawi, R. (2024). Developing evaluative judgement for a time of generative artificial intelligence. *Assessment & Evaluation in Higher Education*.

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & the Teachable Agents Group at Vanderbilt (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, 19(3-4), 363-392.

Chase, C. C., Chin, D. B., Oppezzo, M. A., & Schwartz, D. L. (2009). Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology*, 18(4), 334-352.

Corbin, T., Dawson, P., & Liu, D. (2025). Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education*.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher*, 30(3), 141-158.

Kobayashi, K. (2024). Interactive learning effects of preparing to teach and teaching: A meta-analytic approach. *Educational Psychology Review*.

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, 18(3), 181-208.

Rogers, K., Davis, M., Maharana, M., Etheredge, P., & Chernova, S. (2025). Playing dumb to get smart: Creating and evaluating an LLM-based teachable agent within university computer science classes. *CHI 2025*.

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research*, 77(4), 534-574.

Shahriar, T., & Matsuda, N. (2021). "Can you clarify what you said?": Studying the impact of tutee agents' follow-up questions on tutors' learning. *AIED 2021*, 395-407.

Strömberg, D., Lei, V., & Wu, Y. (2026). The generative AI learning penalty: Evidence from Chinese secondary education. *CEPR Discussion Paper No. DP21577*.

Wang, C., Petrie, C., Stouras, M., et al. (2026). Turning 500+ students into teachers: A semester-long study of an AI teachable agent in an undergraduate algorithms course. *L@S 2026*.