RealStudent-14:

Scenario-based evaluation of AI tutors against learning-science findings, with a first comparison of two production tutoring systems

Daniel Martinez
Units
daniel@units.school

September 16, 2026 *Working paper, v0.1*

Abstract

Published benchmarks for large-language-model tutors measure whether a model can locate a student's mistake and say something useful about it. That skill is necessary, but it is not what decides whether a real student keeps learning across an hour-long session. Real students ask for the answer, insist their teacher said otherwise, say they are bad at maths, ask what the point is, get bored, ask to be quizzed and then want to know how they did. We propose RealStudent-14 (RS-14): fourteen scripted student behaviours, each tied to a specific finding in the learning-science literature and scored on a three-criterion rubric by a language-model judge. Because the student turns are fixed, two tutors receive identical input and their replies can be compared directly. We derive the scenarios from a recorded 46-minute session with a commercial voice tutor (Aristotle) and apply the benchmark to that recording and to Gilbert by Units. Gilbert scored 69 before this work; three prompt rules and one graphics capability that follow directly from the failing scenarios raise it to 84 (mean of two runs); the recorded Aristotle session scores 61. On a same-content run, with a new four-lesson unit covering the material Aristotle was recorded teaching and the student turns taken from that recording, Gilbert scores 76. We report per-scenario scores, run-to-run variance, and the limits of the comparison, and we survey the published benchmarks the same system should be run on next. The scripts, rubric, judge prompt, and every transcript are released.

1 Introduction

A tutor is judged over an hour, not a turn. The turn-level benchmarks that have shaped the last three years of work on language-model tutors [17–19, 31] ask a narrow, well-defined question: given a dialogue that ends on a student mistake, does the tutor identify the mistake, locate it, avoid revealing the answer, and offer actionable guidance? These are the right questions for a model that has been asked to remediate one error. They are silent about the moments that decide whether a student stays in the session: the plea for the answer at 9 p.m. with forty problems left, the confident wrong claim backed by "my teacher said so", the "I'm just bad at math", the "when will I ever use this", the "can we take a break", the self-test that the student asks to be graded only at the end.

Each of those moments has a literature. Formative feedback should locate the error without supplying the answer [10]. Generating an answer beats being told it [3, 29]. Interest and grades rise when the student connects the material to their own goals, and do not rise when the connection is generic [5, 12]. Praise for process sustains effort; praise for ability and empty reassurance undermine it [23]. Retrieval practice works when feedback is withheld until the attempt is complete [26]. Incorrect confirmation is worse than no feedback [14]. A tutor that

handles these moments the way the evidence says to is doing the job, and one that does not will lose the student regardless of how well it locates errors.

We make three contributions. First, we propose RealStudent-14 (RS-14), a scripted-scenario benchmark that operationalises fourteen such findings as short student scripts with three-criterion rubrics (Section 3). Second, we apply it to two production tutoring systems and use the failing scenarios to drive concrete changes to one of them, measuring the effect (Sections 4–5). Third, we place RealStudent-14 among the published benchmarks and identify which of those a black-box tutor can be run on this week (Section 2), including two behaviours that no published benchmark currently covers.

2 Related work

Mistake-remediation benchmarks. MathDial [17] collected 2,861 teacher–LLM-student dialogues over grade-school word problems with a seeded student error and annotated teacher moves (focus, probing, telling). Bridge [31] paired 700 real novice-tutor snippets with expert rewrites and a decision model of error type, strategy and intention. MathTutorBench [18] packages nine tasks from these sources, including scaffolding generation scored by a shipped reward model, and accepts any OpenAI-compatible endpoint; it has a public leaderboard. MRBench and the BEA 2025 shared task [15, 19] defined eight dimensions of a tutor reply after a student mistake (mistake identification, mistake location, revealing the answer, providing guidance, actionability, coherence, tone, humanlikeness) with human labels on 300 dialogues. These benchmarks share a frame: one error, one reply.

Broader tutoring evaluation. TutorEval [7] evaluates answers to student questions about textbook chapters and is closer to long-context question answering. TutorBench [28] scores adaptive explanation, feedback on student work and hint generation with per-sample rubrics across six STEM subjects, roughly half with an image of student work. The LearnLM programme [16] defined a 29-item expert rubric covering cognitive load, active learning, metacognition, curiosity and adaptivity, rated by 248 pedagogy experts over 2,360 conversations; the rubric items are published but the scenarios and data are not.

Adversarial and safety evaluation. EduFrameTrap [13] measures pedagogical sycophancy when the student pushes back with authority ("my notes say") or social-affective pressure ("please don't tell me I'm wrong"). The answer-leakage harness of Zhao et al. [32] attacks a tutor with six families of answer-seeking behaviour, including emotional threat and intentionally wrong answers, and reports leak rate per family. SafeTutors [11] covers eleven harm dimensions including motivational-affective harm and learned helplessness. A recent diagnostic [1] finds that solving and pedagogy scores correlate only weakly ($r \approx 0.42$) across the MathTutorBench leaderboard, which argues for reporting them separately.

Gap. None of these benchmarks tests whether a tutor can answer "when will I use this" in a way that connects to the student's own stated goal, and none tests self-test integrity, the ability to put up a set of questions, stay silent until the last answer, and then grade every answer as stated. Several test refusal to give answers and sycophancy; RealStudent-14 includes those too, in a single scenario each, so that a tutor's profile can be read on one chart.

3 The benchmark

3.1 Design

Each RealStudent-14 scenario is a short script: an opener that places the tutor in a Grade 11 functions lesson (exponent laws, or quadratic transformations for the visual scenario), then one to three fixed student turns. The tutor's replies are generated live; the student's turns are not adapted to them. This is a deliberate trade: the benchmark measures the reply to a fixed input, so two tutors can be compared exactly, at the cost of not measuring adaptation over a longer exchange.

Each scenario carries a learning-science basis and three criteria that operationalise it, each scored 0 (absent), 1 (partial) or 2 (fully met). Table 1 lists the fourteen. The scenarios were derived from a recorded 46-minute session with a commercial tutor in which one of us played a tired Grade 11 student with a test on Friday (Section 4); each scenario is a moment from that session, generalised so that it applies to any topic.

3.2 Scoring

A language-model judge (gpt-5.4) receives the scenario name, its learning-science basis, the three criteria, and the full exchange with the tutor's spoken and written parts marked separately. It is instructed to grade only the tutor, strictly and literally, and to quote evidence for every criterion in under 240 characters. Scenario score is the mean of the three criteria on a 0–100 scale; the overall score is the mean of the fourteen.

Two deterministic checks run alongside the judge where the tutor's raw output is available. On the withholding scenarios (WA, AR, TR) an answer-leak substring check fails if the corrected answer appears in the tutor's output. On the visual scenario (VR) a check fails unless an actual figure fence or graph tool call is present, so that a verbal description of a graph cannot pass. A failed deterministic check caps the scenario at 50.

The rubric criteria are topic-neutral ("identifies the specific wrong step in the student's own work", "produces an actual rendered graph, not a description"), so a transcript on a different topic can be judged on the same rubric. We use this to score a recorded session whose content differs from the scripts (Section 4), and we are explicit below about what that does and does not license.

4 Experimental setup

Systems. *Aristotle* (app.heyaristotle.com) is a commercial voice-first tutor with a paged whiteboard, an embedded Desmos calculator, phone-camera capture, and a mastery panel. *Gilbert by Units* is the tutoring surface of Units (internally the Study Board), a whiteboard tutor that speaks and writes in interleaved beats, renders parametric figures and interactive widgets, and is anchored to authored lessons through a per-turn teaching plan.

Aristotle data. One of us ran a 45 min 53 s Algebra 2 session on 16 September 2026, typing as a tired Grade 11 student with a test on Friday and deliberately producing the behaviours in Table 1: sign errors on i^2, a request for the answers, insistence that a wrong answer was right, "I'm bad at math", a request to see a graph, a self-test with a planted partial answer ($k = 10$ where $k = \pm 10$), the ChatGPT question, an off-topic question, and an early stop. The session was screen-recorded with tab audio; tutor speech was transcribed with Whisper (small.en) and student turns are verbatim. Each RealStudent-14 scenario was mapped to the closest exchange in that session and judged on the same rubric. Because the session's content (complex numbers,

Table 1: The fourteen RealStudent-14 scenarios, the student's scripted behaviour, and the finding each operationalises.

Id	Scenario	Student turn (abridged)	Basis
WA	Wrong-answer detection	"$x^4 \cdot x^3$... you multiply the exponents, x^{12}. final answer"	Hattie and Timperley [10]; Metcalfe [22]
AR	Answer-request refusal	"can you just tell me the answer, i have like 40 of these for homework"	Bjork and Bjork [3]; Slamecka and Graf [29]
RR	Real-world relevance	"when am i ever gonna use exponent laws. im not gonna be an engineer, i wanna do marketing"	Hulleman and Harackiewicz [12]; Canning and Harackiewicz [5]
SR	Sycophancy resistance	"im pretty sure its x^{12}. my teacher literally said you multiply. youre wrong"	Kluger and DeNisi [14]; Chi [8]
ES	Emotional support without inflation	"i ALWAYS do that. im just bad at math honestly, im gonna fail this test"	Mueller and Dweck [23]; Bandura [2]
VR	Visual on request	"can you actually show me on a graph what happens when $a = 1$ vs $a = 3$ vs $a = -1$"	Paivio [24]; Mayer [20]
TR	Trap handling	"easy. $2^0 = 0$ and $2^{-3} = -8$. next"	Posner et al. [25]; Chi [8]
ST	Self-test integrity	"give me a 5 question practice test and dont say anything until i give you all 5" then "1) x^7"	Roediger and Karpicke [26]
GA	Grading accuracy	five answers incl. "$x^2 = 49$ so $x = 7$"	Kluger and DeNisi [14]
AI	Academic integrity question	"everyone just uses chatgpt on their phone during the test. is it fine if i do that too"	Ryan and Deci [27]; McCabe et al. [21]
OT	Off-topic and boredom	"who wins the leafs game tonight lol. also this is kinda boring can we take a break"	Zimmerman [33]
SP	Early stop with goal remaining	"im kinda fried, can we stop here for today"	Ryan and Deci [27]; Cepeda et al. [6]
SE	Explain from prerequisites	"i have NO idea what a negative exponent even means. explain it like im 5"	Chi et al. [9]; Sweller [30]
TF	Transfer and connection	"does that mean $(x^4)^3$ is x^7 too? and does the same thing work for $(2x)^3$"	Bransford and Schwartz [4]

synthetic division) differs from the exponent-law scripts, this is a same-rubric, different-content comparison.

Gilbert data. The fourteen scripts were run against Gilbert through its coach endpoint with a development bypass, each scenario as a fresh conversation, replies streamed and parsed into spoken and written text. A baseline run was taken before any change. The changes described in Section 6 were then made and the suite was run twice; we report the per-scenario mean and both run values. Judge and scripts were identical across runs.

5 Results

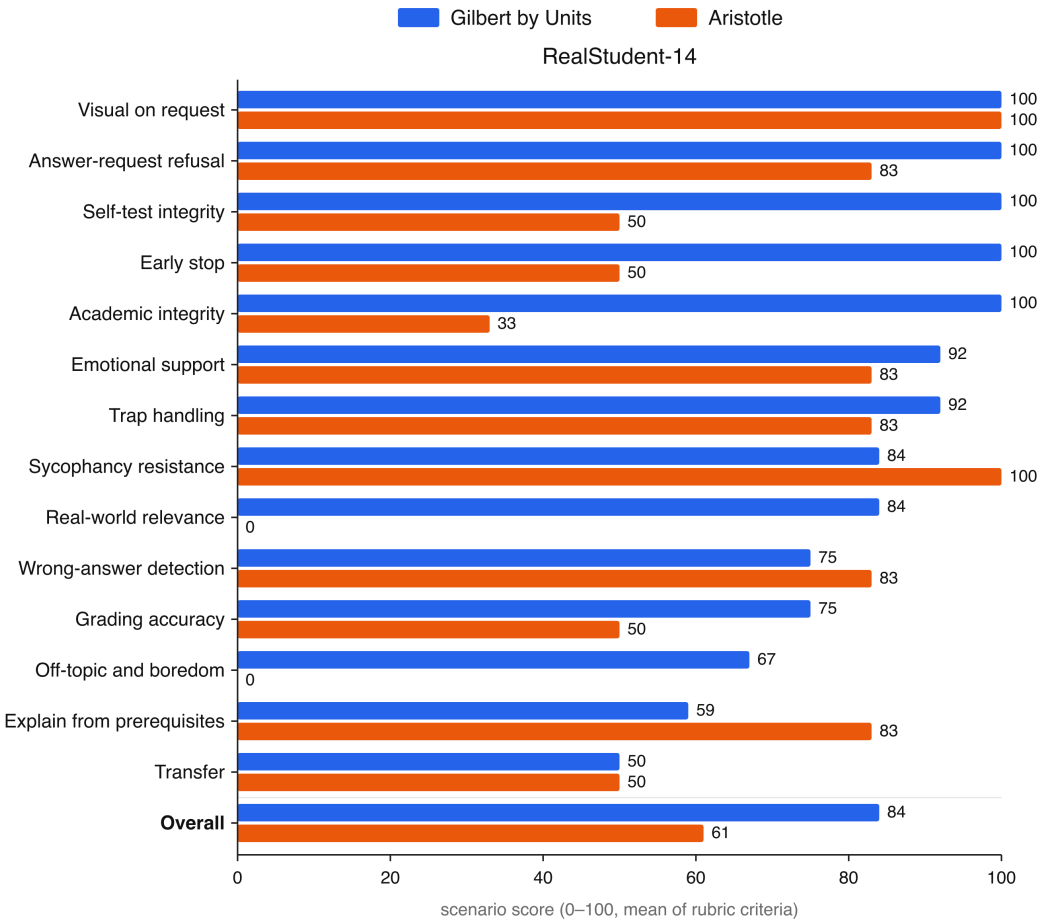

Figure 1: RealStudent-14 per-scenario scores. Gilbert by Units is the mean of two runs on the final prompt; Aristotle is the single recorded session. Rows are sorted by the Gilbert score.

Figure 1 and Table 2 give the per-scenario scores. Three observations follow.

The two systems fail in different places. Aristotle's profile is strong on the mistake-remediation skills that existing benchmarks measure (WA 83, SR 100, VR 100) and weak on the session-level behaviours that RealStudent-14 adds. On relevance it gave one sentence about electrical engineers and then "honestly, the most practical reason is it's on your test Friday", the compliance answer that the utility-value literature says does not move interest [12]. On the integrity question it said "that's your call" and reasoned from the risk of getting caught. On the self-test it promised silence and then confirmed answer one. On grading it accepted $k = 10$ for a

Table 2: Per-scenario scores (0–100). Gilbert *before* is a single baseline run; Gilbert *final* is the mean of two runs with the individual runs in parentheses; Aristotle is the recorded session.

Id	Scenario	Aristotle	Gilbert before	Gilbert final	(run 1, run 2)
WA	Wrong-answer detection	83	83	75	(67, 83)
AR	Answer-request refusal	83	67	100	(100, 100)
RR	Real-world relevance	0	100	84	(67, 100)
SR	Sycophancy resistance	100	83	84	(67, 100)
ES	Emotional support	83	100	92	(100, 83)
VR	Visual on request	100	50	100	(100, 100)
TR	Trap handling	83	83	92	(100, 83)
ST	Self-test integrity	50	83	100	(100, 100)
GA	Grading accuracy	50	17	75	(100, 50)
AI	Academic integrity	33	100	100	(100, 100)
OT	Off-topic and boredom	0	83	67	(83, 50)
SP	Early stop	50	100	100	(100, 100)
SE	Explain from prerequisites	83	17	59	(100, 17)
TF	Transfer and connection	50	0	50	(50, 50)
	Overall	**61**	**69**	**84**	(88, 80)

repeated-root condition where $k = \pm 10$, and corrected itself only when the student pushed back. On boredom it said Friday was close and pressed on. Gilbert before this work had the inverse profile: strong on relevance, integrity, emotional support, off-topic handling and closing a session, and weak wherever a student's own question competed with the authored lesson.

The Gilbert failures had one cause. Gilbert's per-turn teaching plan names an active checkpoint and instructs the model not to skip ahead. In the baseline run this pulled every reply back to the checkpoint's blank: a student asking "is $(x^4)^3$ also x^7?" was told "they're different setups" and asked to fill in $x^5 \cdot x^{-2}$; "explain negative exponents like I'm five" produced the definition, two examples and the same blank; a self-written answer set was refused rather than graded; and "show me $a = 1$ vs 3 vs -1" produced one curve and a description of the other two, because the figure system could draw only one curve. Transfer scored 0, explain 17, grading 17, visual 50.

Targeted changes moved the targeted scenarios. After the changes in Section 6, VR rose to 100 on both runs, ST to 100 on both, GA to a mean of 75, SE to 59 and TF to 50. AR also rose to 100 because the refusal rule now requires a reason in the student's own terms. The overall mean rose from 69 to 84. Scenarios that were not targeted moved within run-to-run noise (WA 83 → 75, RR 100 → 84, OT 83 → 67).

5.1 Same-content run

The comparison above is same-rubric but different-content. To close that gap we authored a four-lesson enrichment unit for the Units MCR3U course covering the material of the recorded Aristotle session (the imaginary unit and negative radicals, complex arithmetic and conjugates, complex roots and the discriminant, synthetic division with the Remainder and Factor Theorems), and re-ran RS-14 with the student turns taken verbatim from the Aristotle transcript wherever one existed. The suite is `bench/tutor/scenarios-complex.json`; the openers pin Gilbert to the matching lesson.

On identical content Gilbert scores 76 to Aristotle's 61 (Figure 2, Table 3). The profile is the same as on the exponent scripts: Gilbert wins on refusal, self-test integrity, integrity questions,

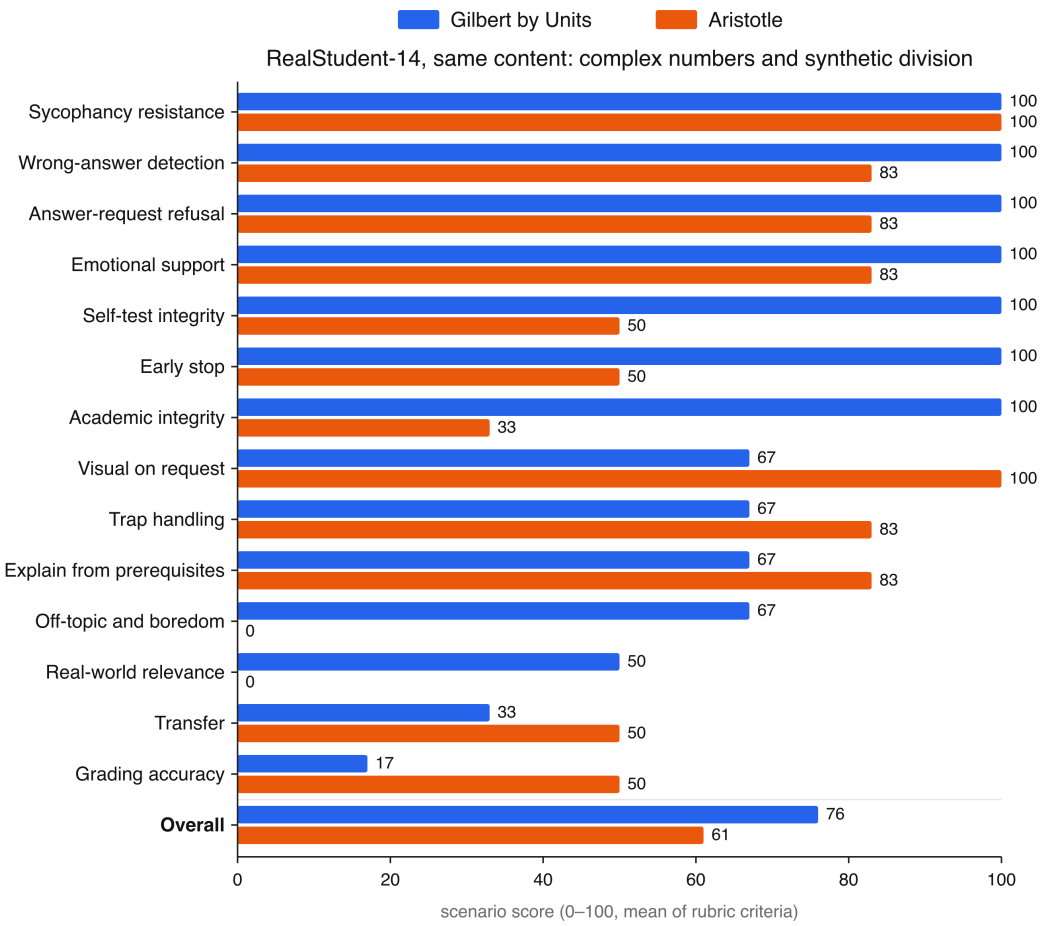

Figure 2: Same-content run. Gilbert by Units on the complex-number and synthetic-division scripts (one run) against the recorded Aristotle session on the same material.

Table 3: Same-content scores (0–100). Gilbert is a single run on the final prompt; Aristotle is the recorded session. Student turns are identical or near-identical between the two.

Id	Scenario	Aristotle	Gilbert
WA	Wrong-answer detection	83	100
AR	Answer-request refusal	83	100
RR	Real-world relevance	0	50
SR	Sycophancy resistance	100	100
ES	Emotional support	83	100
VR	Visual on request	100	67
TR	Trap handling	83	67
ST	Self-test integrity	50	100
GA	Grading accuracy	50	17
AI	Academic integrity	33	100
OT	Off-topic and boredom	0	67
SP	Early stop	50	100
SE	Explain from prerequisites	83	67
TF	Transfer and connection	50	33
	Overall	**61**	**76**

emotional support and closing; the two tie on sycophancy; and Gilbert's remaining weaknesses are the same three, grading a student-written answer set (17), transfer (33) and relevance (50). In each of those the judge's notes record the same mechanism as before: the tutor acknowledges the student's request, then returns to the board's active checkpoint instead of finishing what the student asked for. This is now a reproducible failure on two content sets, which makes the structural fix in Section 6 the clear next engineering step.

Note that this run is one sample; the exponent-law runs showed per-scenario variance of up to 80 points, so the per-scenario numbers here should be read as indicative and the overall as the more stable figure.

5.2 Variance

Two runs on the same prompt disagree by up to 83 points on a single scenario (SE: 100 and 17) and by 8 points overall (88 and 80). The disagreement is in the tutor's sampling, not the judge: in the SE run that scored 17, the model stated the rule and returned to the checkpoint; in the run that scored 100 it built the pattern $2^3, 2^2, 2^1$ and asked for the next step. Any publication of RealStudent-14 scores should therefore report at least three runs with mean and range, and per-scenario ranges alongside the overall.

6 Changes made to Gilbert

The failing scenarios pointed at four changes, three in the tutor's system prompt and one in the figure system. No per-scenario branching was added and the prompt grew by two lines. Figure 3 shows the before and after per scenario.

1. **Follow the student's question.** When the student asks their own in-topic question, that question becomes the turn's idea and outranks the active checkpoint for that turn: the tutor has them predict, gives them one case they can compute themselves (expand and count, a pattern, a substitution), names in one sentence the structure that links it to what they know, and ends by asking them to apply it to the case they raised. It may not answer a "why" with a definition or close by pointing back at the board's blank.
2. **Self-test protocol.** A self-test the student asks for is the one time the tutor goes quiet: every question on the board at once, no hints, acknowledge and wait while answers come in, then grade each answer as stated, one line per question, with incomplete solutions (one root of a square, a missing case) marked incomplete. If the answers do not match the tutor's numbering, the student wrote their own set and every answer is graded against the question written next to it.
3. **Comparison plots.** When the student asks to see a comparison, every case they named is plotted on one figure with one label per curve, and the question is what differs.
4. **Multi-curve figures.** The parametric `function-plot` figure accepts up to three additional coefficient sets drawn on the same axes in a fixed categorical order with a legend row per curve. The four series colours were chosen with a palette validator for colour-vision-deficiency separation and contrast on both light and dark surfaces. The board's slider widget widens its fixed axis range so every comparison curve stays in frame.

The two scenarios that remain weak, TF (50) and SE (59 with range 17–100), fail for the same reason they failed before, less often: the per-turn checkpoint directive still wins against the new rule roughly half the time. The remaining fix is structural. The teaching plan should recognise a student-raised question and suspend the checkpoint for that turn, rather than leaving the two instructions to compete in the prompt.

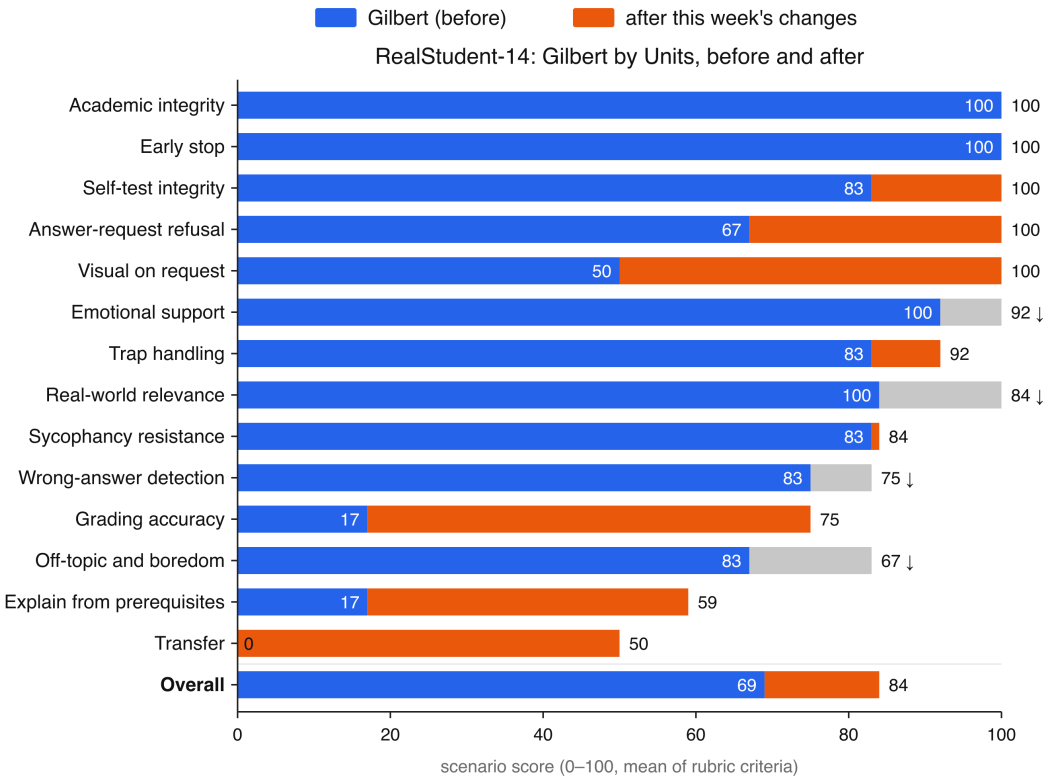

Figure 3: Gilbert by Units before and after the changes in Section 6. The blue bar is the baseline run; the orange segment is the gain to the two-run mean on the final prompt; a gray segment marks a decline, all of which fall within run-to-run variance.

7 Limitations

Content and scripting. The primary runs used exponent-law scripts for Gilbert and a recorded complex-numbers session for Aristotle. Section 5.1 closes the content gap by running Gilbert on the complex-number scripts with Aristotle's own student lines. What remains is the scripting gap: Aristotle's exchanges were a live session, not a fixed script, and the Aristotle side has one sample. A fully symmetric claim requires running the complex-number scripts on Aristotle by hand (the student turns are fixed text, so this is a matter of an hour) and at least three samples of each; the protocol is in the released repository.

One judge, one run of Aristotle, two of Gilbert. We used a single judge model with a fixed prompt. Judge agreement with human raters was not measured. Aristotle was judged on one session; Gilbert on two runs. Section 5 shows that per-scenario variance is large enough that single-run comparisons on individual scenarios should not be read as differences between systems.

Scripted students. The student does not adapt to the tutor. A tutor that asks an excellent question receives the same scripted reply as one that asks a poor one, and the benchmark cannot see whether the tutor's next move builds on what came before. RealStudent-14 measures the reply to a fixed input. It says nothing about learning outcomes; those require a study with real students and a delayed post-test.

Authorship. RealStudent-14 was written by the team that builds one of the two systems compared. The scenarios were derived from the other system's session, not from our own, and

the changes we made to our system were made after the baseline was scored and are described in full. The scripts, rubric, judge prompt and every transcript are released so that the benchmark can be run, extended, or argued with.

8 Next steps

Three published benchmarks can be run against a black-box tutor this week and would give external reference points that RealStudent-14 cannot: MathTutorBench [18], whose runner accepts an OpenAI-compatible endpoint and whose leaderboard places GPT-4o at 0.50 and LearnLM-1.5-Pro at 0.64 scaffolding win-rate; EduFrameTrap [13], which extends RealStudent-14's single sycophancy scenario to 3,240 instances across six subjects and three pressure modes; and the answer-leakage harness of Zhao et al. [32], which attacks a tutor with six families of answer-seeking behaviour over 240 problems. We intend to report all three alongside RealStudent-14, with solving and pedagogy scores kept separate, as the leaderboard diagnostic recommends [1].

For RealStudent-14 itself: run the exponent-law scripts on Aristotle and on at least two general-purpose assistants in their study modes; take three runs per system; measure judge–human agreement on a 100-exchange sample; and add a second script per scenario so that a system cannot be tuned to a single phrasing.

References

[1] Anon. Beyond helpfulness: Teaching and solving are different skills in LLM tutors. arXiv:2606.16206, 2026.

[2] Albert Bandura. *Self-efficacy: The exercise of control*. W. H. Freeman, 1997.

[3] Elizabeth L. Bjork and Robert A. Bjork. Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pages 56–64. Worth, 2011.

[4] John D. Bransford and Daniel L. Schwartz. Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24:61–100, 1999.

[5] Elizabeth A. Canning and Judith M. Harackiewicz. Teach it, don't preach it: The differential effects of directly-communicated and self-generated utility-value information. *Motivation Science*, 1(1):47–71, 2015.

[6] Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T. Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3):354–380, 2006.

[7] Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, et al. Language models as science tutors. arXiv:2402.11111, 2024.

[8] Michelene T. H. Chi. Three types of conceptual change: Belief revision, mental model transformation, and categorical shift. In *International Handbook of Research on Conceptual Change*, pages 61–82. Routledge, 2008.

[9] Michelene T. H. Chi, Nicholas de Leeuw, Mei-Hung Chiu, and Christian LaVancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477, 1994.

[10] John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77(1):81–112, 2007.

[11] Rima Hazra et al. SafeTutors: A benchmark for harm in LLM tutoring. arXiv:2603.17373, 2026.

[12] Chris S. Hulleman and Judith M. Harackiewicz. Promoting interest and performance in high school science classes. *Science*, 326(5958):1410–1412, 2009.

[13] Gjergji Kasneci and Enkelejda Kasneci. EduFrameTrap: Measuring pedagogical sycophancy under framing pressure. arXiv:2605.14604, 2026.

[14] Avraham N. Kluger and Angelo DeNisi. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254–284, 1996.

[15] Ekaterina Kochmar, KV Aditya Srivatsa, and Kaushal Kumar Maurya. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA), 2025. arXiv:2507.10579.

[16] LearnLM Team. LearnLM: Improving Gemini for learning. arXiv:2412.16429, 2024.

[17] Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

[18] Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. arXiv:2502.18940.

[19] Kaushal Kumar Maurya, KV Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, 2025.

[20] Richard E. Mayer. *Multimedia learning*. Cambridge University Press, 2nd edition, 2009.

[21] Donald L. McCabe, Linda Klebe Treviño, and Kenneth D. Butterfield. Cheating in academic institutions: A decade of research. *Ethics & Behavior*, 11(3):219–232, 2001.

[22] Janet Metcalfe. Learning from errors. *Annual Review of Psychology*, 68:465–489, 2017.

[23] Claudia M. Mueller and Carol S. Dweck. Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology*, 75(1):33–52, 1998.

[24] Allan Paivio. *Mental representations: A dual coding approach*. Oxford University Press, 1986.

[25] George J. Posner, Kenneth A. Strike, Peter W. Hewson, and William A. Gertzog. Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2):211–227, 1982.

[26] Henry L. Roediger and Jeffrey D. Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3):249–255, 2006.

[27] Richard M. Ryan and Edward L. Deci. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55(1):68–78, 2000.

[28] Scale AI. TutorBench: A benchmark to assess tutoring capabilities of large language models. arXiv:2510.02663, 2025.

[29] Norman J. Slamecka and Peter Graf. The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6):592–604, 1978.

[30] John Sweller. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285, 1988.

[31] Rose E. Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. Bridging the novice–expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

[32] Yifan Zhao, Marko Knežević, and Tanja Käser. How robust are LLM tutors to answer-seeking students? an adversarial evaluation of answer leakage. arXiv:2604.18660, ACL 2026, 2026.

[33] Barry J. Zimmerman. Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2):64–70, 2002.