# When Socratic Tutoring Scores Track Prompts: Sensitivity Analysis of Simulated Statistics Dialogues

Lauren Mercer
Simon Fraser University
Canada
lauren.mercer@sfu.ca

Maya Kalyani Reed
University of Missouri, Columbia
United States
maya.kalyani.reed@umsystem.edu

Hoang Justin Tran
Université Grenoble Alpes
France
hoang.justin.tran@etu.univ-grenoble-alpes.fr

Alexander J. Garnier
Concordia University
Canada
alexander.j.garnier@concordia.ca

## ABSTRACT

Language-model tutors can generate Socratic-looking dialogue, but evaluation must determine whether questions are correct, learner-responsive, and educationally meaningful. This study reanalyzes a completed prompt comparison using Qwen3-14B as both tutor and simulated learner. Four tutor conditions—direct explanation, minimal Socratic instruction, fixed Socratic script, and adaptive Socratic prompt—were crossed with three introductory-statistics topics, two prompt phrasings, and three persona-linked seeds, yielding 72 eight-turn dialogues. A mechanically scored Open Socratic Index initially gave the adaptive condition a mean of 63.0 and the fixed script a mean of 62.9, a paired difference of 0.1 points. The adaptive condition appeared to exceed direct explanation by 8.93 points, but 8.67 points came from a transfer detector that rewarded wording mandated by the adaptive treatment. Removing that detector reduced the contrast to 0.26 points, with an exact sign-flip p-value of 0.935. The protocol also gave tutors full numerical context while withholding it from simulated learners, so content accuracy could not be compared fairly across conditions. In this corpus, prompt rankings reflected information access and metric construction rather than demonstrated pedagogical adaptivity. The contribution is a diagnostic workflow for decomposing composite tutoring scores, testing treatment-independent sensitivity, and specifying evidence needed before simulated dialogues support educational claims.

## KEYWORDS

AI education, assessment, scholarly integrity

## 1 INTRODUCTION

Learning technologies that use language models increasingly promise low-cost dialogue, questioning, and feedback for domains such as introductory statistics, where learners often struggle with interpretation, uncertainty, evidence, and causal reasoning. However, a dialogue can look Socratic while giving incorrect guidance, ignoring the learner's state, or following a script that happens to match an automatic rubric. For researchers who build and evaluate such systems, the central problem is not only whether a tutor asks questions, but whether evaluation separates surface form from responsive statistical instruction. This study examines that problem through a controlled prompt experiment with one open-weight model and treats the composite Socratic score as an object of validation. It contributes a diagnostic analysis of score components, treatment-coupled detectors, information-access asymmetries, and evidence requirements for future evaluations of simulated tutoring dialogues.

Large language models (LLMs) can sustain extended educational dialogue, ask questions, and produce feedback at low marginal cost. Those capabilities make them attractive for statistics instruction, where common difficulties concern interpretation, uncertainty, evidence, and causal reasoning rather than computation alone. Yet a dialogue that looks Socratic is not necessarily correct, responsive, or educationally useful. A tutor can ask many questions while withholding essential information, reward a fluent misconception, or follow a predetermined script without representing the learner's state. Evaluation must therefore separate the form of Socratic dialogue from the substance of statistical guidance.

Vilakati recently reported that simulated LLM tutoring dialogues could exhibit inquiry, knowledge construction, dialectical examination, reflective transfer, and learning progression under a Socratic Learning Analysis Scheme [36]. That result motivates a narrower field question: when one open model is evaluated under controlled prompt conditions, do apparent gains reflect adaptive guidance, or do they reflect the prompt and the measurement procedure? The distinction matters because a measure can become coupled to a treatment. If an adaptive prompt explicitly requests a general rule and a transfer example, then a detector that rewards the phrases "general rule" or "another example" partly measures instruction compliance rather than an independently established educational construct.

Dialogic pedagogy provides a demanding standard. Productive dialogue builds on learner contributions, exposes reasons, tests alternatives, supports revision, and connects an idea to a new situation [1, 14, 15, 26, 39]. Intelligent tutoring systems similarly distinguish domain knowledge, learner state, pedagogical policy, and surface realization [12, 30]. These traditions imply that a valid evaluation should give competing conditions equivalent information, assess correctness, and test whether the tutor's next action changes with the learner's state.

Related work in machine learning shows how carefully designed tasks, representations, and measurements can support targeted evaluation. For example, semi-supervised segmentation, in-context semi-supervised learning, novel-object captioning, referring-expression grounding, model reprogramming, and generative biomedical reconstruction each use task-specific metrics to probe particular capabilities [6–9, 19, 37, 40, 41]. These domains are technically distant from statistics tutoring, but they illustrate that evaluation design matters: a metric must reflect the construct of interest rather than the easiest signal in the data. In educational dialogue, automatic evaluation faces an additional difficulty because the generated interaction changes the evidence available to later turns. A small prompt difference can alter both what a learner says and what an automatic metric subsequently rewards.

This study examines that difficulty through a completed factorial experiment with one open-weight model. Four tutor prompts are crossed with three statistics topics, two instruction phrasings, and three persona-linked seeds. The original analysis used a transparent composite called the Open Socratic Index. The present paper treats that score as an object of validation rather than as a validated endpoint. We decompose the score, inspect information access, locate treatment-coupled features, and rerun the paired comparisons after removing the transfer component.

The study makes four contributions:

- It provides a fully specified four-condition prompt comparison across multiple topics, phrasings, and stochastic runs.
- It introduces a component-contribution analysis that shows how each detector affects a composite tutoring score.
- It demonstrates a treatment-independence sensitivity test in which prompt rankings are recomputed after removing a detector directly solicited by the treatment.
- It formulates protocol requirements for future evaluations: context parity, correctness scoring, explicit learner-state annotations, authentic-response validation, and uncertainty models aligned with the intended population.

The findings do not establish educational effectiveness. Instead, they identify why the current experiment cannot support such a conclusion and show how a superficially favorable prompt comparison can collapse under elementary validity checks.

## 2 RELATED WORK

### 2.1 Socratic and dialogic instruction

Socratic instruction uses questions to elicit a claim, expose its supporting reasons, examine alternatives, and encourage revision. Early computational formulations represented tutoring as goals and inquiry moves, while later systems tested whether learners could apply ideas beyond the immediate exchange [33, 38]. Classroom studies likewise associate productive questioning with the examination of assumptions and evidence [17]. For statistics education, Amin describes a student-centered approach organized around meaningful problems and disciplined questioning [3].

Dialogic teaching broadens the focus from the number of questions to the quality of participation. Coding schemes distinguish cumulative, exploratory, authoritative, and disputational talk, and their interpretation depends on the unit of analysis and the context in which an utterance occurs [14–16]. Dialogic quality also concerns whether a learner's contribution changes the trajectory of the conversation [1, 2]. This emphasis is important for LLM tutors because question marks and reasoning words are easy to elicit through prompting, whereas contingent guidance requires the system to interpret what the learner has actually said.

### 2.2 Language-model tutoring and adaptivity

Language-model tutoring research increasingly combines generation with explicit pedagogical structure. MWPTutor constrains responses through a handcrafted instructional design and guardrails [28]. Chained tutoring separates interaction, reflection, and reaction into components that can be evaluated independently [5]. Other systems use prompting or fine-tuning to elicit critical-thinking moves [10] and examine the contribution of symbolic or contextual scaffolding [11]. These approaches inherit a longstanding intelligent-tutoring principle: pedagogical control should be observable separately from fluent response generation.

The closest work sets a substantially higher bar than surface-form scoring. SocraticLM trains an open Socratic teaching model using simulated cognitive states and evaluates several teaching abilities [20]. MathTutorBench organizes evaluation around multiple teacher skills and tasks, with teacher-grounded targets and a learned reward model for open-ended responses [25]. GuideEval uses authentic educational dialogues to evaluate learner-state perception, instructional orchestration, and elicitation as separate capacities [21]. Borchers and Shou directly test whether tutors adapt to differing error and knowledge states rather than merely follow a pedagogical style prompt [4]. Together, these studies show that adaptivity is a relation between learner state and tutor action, not a property established by naming a prompt "adaptive."

Learning outcomes remain a separate evidential layer. In an algebra-hint study, LLM-generated hints that passed quality screens did not automatically reproduce the effects of human-authored hints [29]. Meta-analyses of intelligent tutoring systems estimate learning effects from comparative student outcomes and document substantial heterogeneity across systems and contexts [23, 32, 35]. Dialogue-form measures can help screen a system, but they cannot substitute for immediate learning, delayed retention, or transfer outcomes.

### 2.3 Simulated learners and automatic measurement

Simulated learners are useful for early system development because they enable repeatable stress tests without exposing students to an unvalidated tutor. Generative Students, SOEI, MathVC, and related work investigate persona-conditioned or multi-agent learner simulation [22, 24, 43]. Their scientific use depends on fidelity. A capable model instructed to behave like a novice may retain expert knowledge, revise because of conversational pressure, or produce linguistically plausible states that do not match human learning processes [18, 42].

Recent evidence strengthens this concern. Scarlatos and colleagues benchmark student-simulation methods against authentic tutoring dialogues using linguistic, behavioral, and cognitive criteria and find that prompted simulations do not reliably substitute for real

student responses [31]. EvalConvoLearn likewise grounds simulation assessment in authentic tutoring distributions [27]. A simulated learner can therefore support software testing, but its responses require validation before they are treated as a population of students.

Automatic dialogue scoring creates a parallel risk. Generative models can classify tutoring dialogue acts, but performance depends on the codebook, context, and validation target [13]. Educational measurement guidance calls for documented construct development, disagreement analysis, and external validity evidence rather than treating reproducibility alone as validity [34]. A deterministic metric can be perfectly reproducible and still reward the wrong behavior.

Vilakati's preprint supplies the motivating statistics problem, simulated tutor–learner setup, and emphasis on inquiry, construction, examination, transfer, and progression [36]. The present study adds direct, minimal, fixed-script, and adaptive conditions with an open model, then evaluates whether information access and score construction permit those conditions to be compared.

# 3 METHOD

## 3.1 Research questions and scope

We study prompt-policy evaluation under a fixed eight-turn, model-to-model simulation with one open-weight instruction model. The study addresses three questions:

(1) How do direct, minimal Socratic, fixed-script, and adaptive prompts differ on the original Open Socratic Index?
(2) Which score components account for the observed condition contrasts?
(3) Do the contrasts remain after removing a treatment-coupled transfer detector?

The intended unit is the generated dialogue cell, defined by topic, prompt phrasing, and seed. The experiment does not sample human learners, schools, instructors, or model families. Consequently, its numerical summaries describe this generated corpus and do not estimate effects for educational populations.

## 3.2 Factorial design

The design crosses four tutor conditions, three topics, two prompt phrasings, and three seeds, producing 72 dialogues. Each dialogue contains eight tutor–learner turn pairs, for 1,152 model completions.

The direct condition explains the correction and asks a comprehension question. The minimal condition instructs the model to use a Socratic style without prescribing a sequence. The fixed-script condition requires six moves in order: elicit a claim, request evidence, request a calculation or comparison, introduce a counterexample, request revision, and request transfer. Because conversations last eight turns, the script does not define the final two tutor actions. The adaptive condition asks the tutor to choose one move from clarification, evidence, calculation, counterexample, hint, revision, or transfer based on the latest learner response. It also requires the ending to elicit a general rule and a transfer example.

The three topics are chi-square independence, confidence-interval interpretation, and confounding. Each topic starts from an author-written misconception. The prompt phrasing is either canonical or a semantic paraphrase. Seeds 1103, 2207, and 3301 are deterministically associated with cooperative, hesitant, and initially confident personas; seed effects and persona effects are therefore not separately identifiable.

## 3.3 Information presented to the two roles

Tutor and learner histories are separate. The tutor receives the full problem context, the learner's starting misconception, the learning target, and the condition-specific policy. The learner receives a persona, the starting misconception, and general response instructions. The learner does not receive the full numerical problem context in its system history.

This information imbalance is decisive for content evaluation. In the chi-square task, Socratic conditions can ask the learner to calculate proportions before the learner has seen the group totals. The learner may then invent a denominator, after which the tutor may correct or endorse it. Direct tutoring often reveals the totals while explaining the solution, so the conditions do not expose the simulated learner to equivalent evidence. Differences in correctness, revision, or apparent responsiveness cannot be separated from differences in information disclosure.

The issue is not resolved by separate histories. Separate histories prevent the learner from reading the tutor's system prompt, but they do not ensure that both conditions begin with equivalent task knowledge. A valid rerun must provide the same complete problem statement to every learner trajectory before the first condition-dependent tutor action.

## 3.4 Model and generation

Qwen3-14B serves both roles through a local OpenAI-compatible endpoint. Sampling uses temperature 0.7, top-p 0.8, and a maximum of 256 output tokens. Turn-specific offsets are added to the three dialogue seeds, and thinking output is disabled in the serving template. Requests, responses, token counts, and wall times are stored in the raw JSON Lines corpus.

The manifest names the served weight snapshot as `Qwen/Qwen3-14B snapshot 40c069824f4251a91eefaf281ebe4c544efd3e18`. Its reported SHA-256 value is the hash of that description string, not a hash of the model files. The record also omits the inference-server version, tokenizer and chat-template revision, numerical precision or quantization, and hardware. The raw responses identify a vLLM system fingerprint, but the available artifacts are not sufficient to recreate the complete serving environment.

## 3.5 Open Socratic Index

The Open Socratic Index (OSI) is a deterministic score from 0 to 100. It combines question rate, one-question discipline, reasoning-probe rate, lexical response-contingent feedback, learner revision, transfer, and autonomous reasoning. The respective weights are 0.18, 0.10, 0.18, 0.14, 0.16, 0.12, and 0.12.

Question rate checks whether a tutor turn contains a question mark. One-question discipline checks whether it contains exactly one question mark. Reasoning probes search for terms such as

"why," "how," "evidence," "reason," and "compare." Responsive feedback searches tutor turns for phrases such as "you said," "your," or "building on." Revision searches the final learner turns for generic revision phrases or topic terms. Autonomous reasoning searches the final learner turns for connectors such as "because," "so," or "depends," while excluding a small set of misconception phrases.

Transfer is marked positive if either tutor or learner text contains one of several expressions: another or new example, study, situation, or context; "elsewhere"; "transfer"; "general rule"; or "suppose instead." This component has two validity problems. First, the adaptive condition explicitly instructs the tutor to end with a general rule and transfer example, so the detector rewards mandated treatment language. Second, "elsewhere" is a literal category name in the chi-square problem, which can trigger the detector without a new task or context.

OSI includes no factual-correctness component. A dialogue can therefore receive credit for questioning, revision, or transfer even when its quantities or statistical interpretation are wrong. The score was not calibrated against blinded expert ratings, developed on a separate corpus, or locked before the reported transcripts were examined. Its values are descriptive properties of these regexes on this corpus.

### 3.6 Analysis

The original analysis reports condition means and nonparametric bootstrap intervals that resample the 18 dialogue cells within each condition. Paired contrasts match topic, phrasing, and seed, then use a randomized sign-flip test. Those intervals and tests are retained as descriptions of the fixed corpus, but they do not represent uncertainty over unseen topics, students, models, or prompts. The three topics and two near-paraphrases are author-selected, while each seed is confounded with a persona.

We add two sensitivity analyses. First, for every OSI component, we compute its weighted contribution to the adaptive-minus-direct mean difference. Second, we remove the 12-point transfer component from every dialogue and recompute paired mean differences. Exact two-sided sign-flip p-values enumerate all sign assignments across the 18 matched cells. The sensitivity analysis uses the stored dialogue-level metrics and does not regenerate any dialogue.

We also identify where transfer matches occur by applying the original transfer regexes separately to tutor and learner text. A tutor-only match shows that the metric can grant transfer credit without evidence that the learner produced or applied a generalization.

## 4 RESULTS

### 4.1 Descriptive OSI outcomes

All four conditions contain 18 completed dialogues. Mean OSI values are 54.1 for direct explanation, 58.7 for the minimal Socratic prompt, 62.9 for the fixed script, and 63.0 for the adaptive prompt. The adaptive bootstrap interval is 58.0–67.5. These values describe regex matches and should not be interpreted as estimates of tutoring quality.

The original paired analysis gives adaptive-minus-direct as 8.93 points with $p = 0.0287$, adaptive-minus-minimal as 4.33 points with $p = 0.2207$, and adaptive-minus-scripted as 0.07 points with $p = 0.9876$. The last contrast does not test equivalence because no equivalence margin was specified and the design does not isolate adaptivity. It only shows that the two prompt conditions receive nearly identical aggregate scores under OSI.

### 4.2 Component decomposition

The adaptive-minus-direct difference is not distributed evenly across the index. Transfer contributes 8.67 of the 8.93 points. Reasoning probes, feedback, and revision contribute 1.75, 1.44, and 1.78 points, respectively. Adaptive prompting is 2.71 points lower on one-question discipline and 2.00 points lower on autonomous reasoning, while question rate contributes no difference.

Figure 1 shows why the aggregate is misleading. The adaptive prompt directly requests a transfer ending, and the transfer regex rewards "general rule," "another example," and related phrases. The largest positive component is therefore also the component most clearly coupled to the treatment instruction.

The source check further weakens the transfer interpretation. Five fixed-script dialogues and five adaptive dialogues receive transfer credit from tutor text without any transfer-regex match in learner text. A tutor can thus satisfy the metric by asking for transfer even when the learner does not supply it. In the chi-square task, the category label "elsewhere tattoo" can also count as transfer despite remaining within the original problem.

### 4.3 Sensitivity without the transfer detector

Removing the transfer component changes the ranking of the prompt conditions. The adaptive-minus-direct difference becomes 0.26 points with exact $p = 0.935$. Adaptive-minus-minimal becomes -2.33 points with exact $p = 0.349$, and adaptive-minus-scripted becomes -3.93 points with exact $p = 0.232$.

Figure 2 demonstrates that the apparent adaptive advantage is not robust to an elementary construct check. This does not establish that direct instruction is better. It establishes that the available metric cannot support a stable ranking because one detector both receives substantial weight and overlaps with the defining language of the treatment.

### 4.4 Information imbalance and correctness

The information protocol prevents a clean content comparison. The learner starts with a misconception but not the numerical context needed to answer early calculation questions. Transcript inspection shows invented denominators and other fabricated quantities in chi-square dialogues. For example, some learner turns treat the no-tattoo group as size 20 or 100 before being corrected, and some adaptive transcripts temporarily describe the parlor or "elsewhere tattoo" groups as if they contained 200 or 300 people. In several runs, a tutor turn builds on the fabricated value before later turns repair it. Because direct explanations reveal more of the table than question-led conditions, factual failure is partly a consequence of condition-dependent disclosure.

OSI does not register these errors. A learner can use a revision phrase after an incorrect exchange, and a tutor can receive transfer credit while repeating treatment language. The absence of correctness scoring is especially serious in statistics education, where an

**Table 1: Descriptive Open Socratic Index (OSI) outcomes. Intervals resample the fixed dialogue cells and do not describe an educational population.**

| Condition | Dialogues | Mean OSI | Bootstrap interval | Question rate | Revision rate | Transfer rate |
|---|---|---|---|---|---|---|
| Direct | 18 | 54.1 | 49.2–58.8 | 100.0% | 66.7% | 16.7% |
| Minimal Socratic | 18 | 58.7 | 53.5–63.3 | 100.0% | 66.7% | 33.3% |
| Fixed script | 18 | 62.9 | 58.1–67.4 | 97.2% | 72.2% | 55.6% |
| Adaptive Socratic | 18 | 63.0 | 58.0–67.5 | 100.0% | 77.8% | 88.9% |

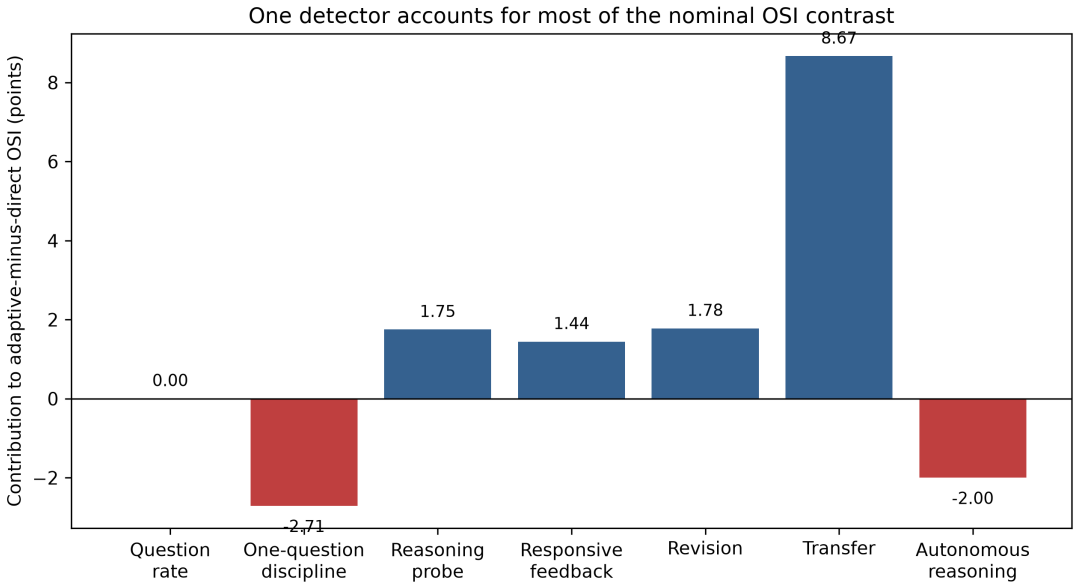


**Figure 1: Weighted contribution of each OSI component to the adaptive-minus-direct mean. Positive bars favor the adaptive prompt and negative bars favor direct explanation. Transfer supplies 8.67 of the 8.93 reported points, while one-question discipline and autonomous reasoning favor direct explanation.**

**Table 2: Sensitivity of the adaptive contrasts to removal of the transfer detector. Exact two-sided sign-flip tests enumerate all assignments.**

| Baseline | Reported adaptive-minus-baseline OSI | Difference without transfer | Exact sign-flip p-value without transfer | Paired cells |
|---|---|---|---|---|
| Direct | 8.93 | 0.26 | 0.935 | 18 |
| Minimal Socratic | 4.33 | -2.33 | 0.349 | 18 |
| Fixed script | 0.07 | -3.93 | 0.232 | 18 |

interaction can be conversationally fluent yet wrong about denominators, uncertainty, significance, or causation.

No condition-level factual-error rates are reported as primary outcomes because the existing transcripts lack blinded, locked annotations, and we do not re-label the corpus after seeing the results. The available inspection is sufficient to invalidate a pedagogical ranking but not to replace the score with a post hoc accuracy leaderboard.

## 4.5 Prompt wording, topics, and seeds

The corpus contains visible variability across prompt phrasings, topics, and persona-linked seeds. For direct tutoring, the canonical and paraphrased means are 48.6 and 59.5, a shift of 10.9 points. The adaptive means are 61.8 and 64.1. These values show that a single prompt form would give an incomplete description of the system.

The design nevertheless cannot support a general robustness claim. There are only three author-written topics and two closely

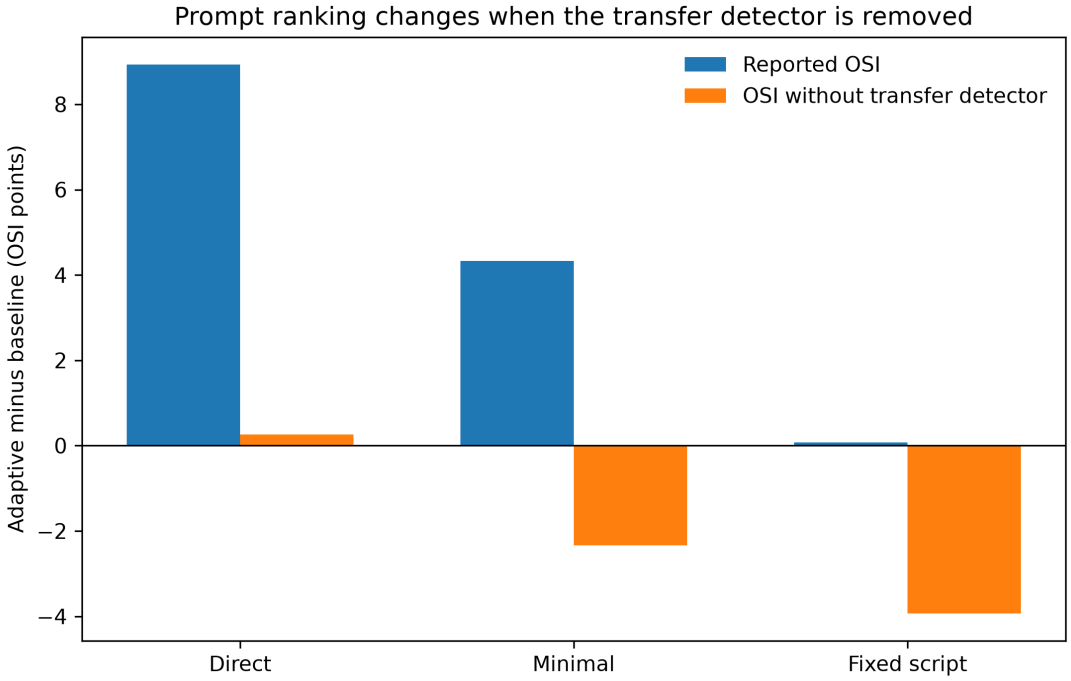


**Figure 2: Adaptive-minus-baseline contrasts under the reported OSI and after deleting the transfer detector. Removing one treatment-coupled component reduces the direct contrast from 8.93 to 0.26 points and reverses the signs of the minimal and fixed-script contrasts.**

related phrasings. The same three seeds always select the same personas, and matching seed identifiers across tutor conditions do not create matched human learning trajectories. A future study needs independent variation in item, learner, prompt, model, and rater, with those sources represented in the analysis.

## 5 DISCUSSION

### 5.1 What the experiment establishes

The experiment establishes that Qwen3-14B can produce different surface patterns under different tutoring prompts. Structured prompts increase the frequency of regex-detected transfer language, and all conditions generate many question-bearing tutor turns. The raw artifacts also show that prompt wording and stochastic generation affect the resulting dialogues.

The experiment does not establish that the adaptive prompt is a better tutor. The adaptive and fixed-script means are nearly identical, and adaptivity is neither isolated nor directly measured. More importantly, the nominal adaptive-over-direct contrast is almost entirely attributable to a detector that rewards wording demanded by the adaptive instruction. Once that component is removed, the contrast is close to zero.

This distinction is central to evaluation design. Prompt compliance can be useful during development, but it is not the same construct as learner-responsive pedagogy. A system can follow an instruction to ask for a transfer example without recognizing whether the learner is ready for transfer. It can mention the learner's words without accurately diagnosing the misconception. It can produce a final revision without causing conceptual change.

### 5.2 Why adaptivity remains unresolved

An adaptive tutor should map distinct learner states to appropriately distinct instructional actions. Demonstrating that relation requires at least three forms of evidence. First, the study needs learner-state labels or other defensible state representations. Second, tutor moves must be annotated at a level that distinguishes clarification, diagnosis, hinting, explanation, counterexample, and transfer. Third, matched counterfactual states should test whether changing the learner response changes the selected move while holding the problem and history constant.

The current prompts do not provide this test. Fixed script and adaptive instruction differ in sequence, allowed actions, hinting, answer-disclosure policy, and ending requirements. The fixed script specifies six moves for an eight-turn conversation, leaving two actions unconstrained. A difference between the conditions would therefore mix several prompt features, while a null difference would not show that adaptivity lacks value. GuideEval's separation of perception, orchestration, and elicitation provides a more appropriate structure for future evaluation [21].

### 5.3 Measurement validity before composite scores

Composite measures are attractive because they compress a long dialogue into one value. Compression is useful only when the components measure defensible constructs. OSI's transparent implementation makes its failure mode visible: lexical features can be traced to their weights and prompt instructions. That transparency is valuable, but it does not confer validity.

A stronger measure should begin with a codebook developed independently of the evaluated conditions. Blinded statistics educators should annotate factual correctness, misconception diagnosis, pedagogical action, response contingency, answer leakage, learner revision, and genuine transfer. Multiple raters are needed because these judgments are not reducible to punctuation or keyword counts. Disagreement should be reported at the item level, and any automatic scorer should be trained or tuned on a development split and frozen before held-out evaluation.

Correctness should be a gating dimension rather than an optional component. A statistically incorrect dialogue should not receive a high pedagogical score merely because it contains questions and revision language. Transfer should require learner application to a genuinely new problem, not the tutor's request for an example or the appearance of a category word. Revision should require a correct change from a documented prior state, and response contingency should require a substantively appropriate action rather than lexical reference to "your" answer.

### 5.4 Appropriate use of simulated learners

Simulated learners can support controlled engineering tests. They can expose a tutor to hesitation, confident misconception, terse answers, or repeated errors without risk to students. They are also useful for debugging turn management, checking answer leakage, and generating candidate failure cases for human review.

They should not be treated as interchangeable with students. The same model family plays tutor and learner here, creating shared pretraining, style, and solution knowledge. Persona prompts do not remove that knowledge. The learner may revise because the tutor is authoritative, not because an explanation changed a human mental model. It may also generate facts that were never supplied, as the missing-context problem illustrates.

Future simulator use should be grounded in authentic response distributions. One approach is to condition on real student histories and evaluate next-response generation against held-out human turns. Another is to compare simulator behavior across independent model families and test whether tutor rankings are stable. If rankings change with the simulator, the simulator should be treated as an experimental factor rather than a neutral stand-in for learners.

### 5.5 Statistical interpretation

The reported bootstrap intervals resample 18 fixed dialogue cells per condition. They do not describe uncertainty over a population of statistics problems, students, or models. Similarly, the paired sign-flip tests operate on cells whose seeds are tied to personas and whose trajectories diverge after the first tutor action. The resulting p-values are descriptive sensitivity statistics, not evidence of a general educational effect.

The p-value of 0.9876 for adaptive versus script also does not establish equivalence. Equivalence requires a justified margin and an analysis designed to determine whether an interval lies within that margin. With three tasks and an unvalidated outcome, no defensible educational equivalence margin is available.

A future confirmatory analysis should define the target population before selecting a resampling or hierarchical model. If the target concerns unseen topics, learners, prompts, tutor models, and raters, each source must vary independently and enter the model. Primary outcomes and sensitivity tests should be preregistered, and multiplicity should be addressed when several prompt contrasts or component outcomes are tested.

### 5.6 Implications for statistics tutoring research

Statistics tutoring is a particularly revealing domain for these evaluation problems. A response may sound reasonable while confusing raw counts with proportions, probability with confidence coverage, association with causation, or statistical significance with practical importance. Those errors are not cosmetic; they are the content being taught.

A useful evaluation pipeline should therefore place vetted task context outside the condition-specific tutor policy and make it equally available to every learner. It should record the tutor's inferred learner state and chosen pedagogical action. It should score correctness and calibration at each turn, then evaluate whether the learner produces a correct explanation on a held-out transfer task. Only after those conditions are met is it meaningful to ask whether a prompt, script, or learned policy improves dialogue quality.

Human evaluation remains necessary even when automatic measures are used. Experienced statistics educators can distinguish a productive question from a leading one, a useful hint from answer disclosure, and genuine conceptual transfer from paraphrase. Human learner studies are then needed to determine whether the measured dialogue behaviors improve immediate understanding, retention, or transfer.

## 6 LIMITATIONS

This study uses one open-weight model in both roles and a small set of author-written problems. It includes no authentic student responses, human learners, delayed outcomes, or instructor judgments. The learner personas are not calibrated cognitive models, and the three seeds are confounded with those personas.

The information available to tutor and learner is unequal. This flaw affects every content-sensitive comparison in the corpus. The OSI is unvalidated, contains no correctness dimension, and includes lexical detectors that overlap with treatment instructions. Its bootstrap intervals do not generalize beyond the fixed generated cells.

Adaptivity is not isolated because the scripted and adaptive conditions differ along several prompt dimensions, and the fixed sequence does not specify all tutor turns. No learner-state labels, policy traces, counterfactual state tests, or preferred-action annotations are available. The study therefore cannot determine whether the adaptive prompt selected actions in response to learner state.

Transcript inspection was not conducted with a locked codebook or blinded independent raters. It identifies sufficient factual failures to reject the current ranking, but it does not provide reliable condition-level error estimates. The serving manifest also lacks immutable model-file hashes and several environment details required for exact regeneration.

These limitations make the current evidence unsuitable for a claim about tutoring effectiveness, educational adaptivity, or deployment readiness. They also limit the work's contribution relative to established benchmarks and model-based tutors that use broader tasks, authentic dialogue, state-aware evaluation, or validated human judgments [20, 21, 25, 31].

## 7 CONCLUSION

Prompted Socratic tutoring cannot be evaluated reliably by counting Socratic-looking language alone. In this corpus, adaptive and fixed-script prompting receive almost identical aggregate scores, while the nominal adaptive advantage over direct explanation is dominated by a transfer detector that overlaps with the adaptive instruction. Removing that detector changes the adaptive-minus-direct contrast from 8.93 to 0.26 points.

The protocol also withholds full problem context from the simulated learner, allowing information disclosure to differ by tutor condition. Because the score does not measure correctness, fluent but factually compromised exchanges can still receive high values. These failures prevent the current experiment from ranking tutoring policies.

The practical result is a set of design requirements for the next study: equal task information, correctness-sensitive and treatment-independent outcomes, authentic-response validation, direct tests of learner-state adaptivity, independent variation across tasks and learners, and blinded expert annotation. With those elements in place, simulated dialogue can serve as a useful development instrument. Without them, composite prompt scores risk measuring the instructions and detectors more than the quality of teaching.

## DATA AVAILABILITY

The source contingency-table context is available in the public preprint at https://doi.org/10.20944/preprints202610.0576.v1 under a CC BY 4.0 license. The two additional statistics vignettes, prompts, 72 generated dialogues, derived metrics, and sensitivity outputs are included with the manuscript artifacts.

## CODE AVAILABILITY

Code is available from the corresponding author and will be released at publication.

## CREDIT AUTHOR CONTRIBUTION

Lauren Mercer, Maya Kalyani Reed, Hoang Justin Tran, Alexander J. Garnier: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—original draft, Writing—review and editing, Visualization, Project administration.

## FUNDING

This research received no external funding.

## DECLARATION OF COMPETING INTEREST

The author declares no competing interests.

## DECLARATION OF GENERATIVE AI USE

During the preparation of this work the author used AI-assisted tools for literature search, code development and drafting. The author reviewed and edited the content and takes full responsibility for it.

## REFERENCES

[1] Robin Alexander. 2018. Developing dialogic teaching: genesis, process, trial. *Research Papers in Education* 33, 5 (June 2018), 561–598. https://doi.org/10.1080/02671522.2018.1481140

[2] Robin Alexander. 2019. Whose discourse? Dialogic Pedagogy for a post-truth world. *Dialogic Pedagogy: An International Online Journal* 7 (Mar 2019). https://doi.org/10.5195/dpj.2019.268

[3] Zeinab Amin. 2026. Socratic Method: A Student‚ÄêCentered Approach to Teaching Statistics. *Teaching Statistics* (June 2026). https://doi.org/10.1002/test.70044

[4] Conrad Borchers and Tianze Shou. 2025. Can Large Language Models Match Tutoring System Adaptivity? A Benchmarking Study. *arXiv preprint arXiv:2504.05570* (2025). arXiv:2504.05570 https://arxiv.org/abs/2504.05570

[5] Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Empowering Private Tutoring by Chaining Large Language Models. *arXiv preprint arXiv:2309.08112* (2023). arXiv:2309.08112 https://arxiv.org/abs/2309.08112

[6] Ming-Yu Chung, Jiashuo Fan, Hancheng Ye, Qinsi Wang, Wei-Chen Shen, Chia-Mu Yu, Pin-Yu Chen, and Sy-Yen Kuo. 2025. Model Reprogramming Demystified: A Neural Tangent Kernel Perspective. *arXiv preprint arXiv:2506.00620* (2025). arXiv:2506.00620 https://arxiv.org/abs/2506.00620

[7] Jiashuo Fan, Bin Gao, Huan Jin, and Lihui Jiang. 2022. UCC: Uncertainty guided Cross-head Cotraining for Semi-Supervised Semantic Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 9937–9946. https://doi.org/10.1109/cvpr52688.2022.00971

[8] Jiashuo Fan, Yaoyuan Liang, Leyao Liu, Shaolun Huang, and Lei Zhang. 2023. RCA-NOC: Relative Contrastive Alignment for Novel Object Captioning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 15464–15474. https://doi.org/10.1109/iccv51070.2023.01422

[9] Jiashuo Fan, Paul Rosu, Aaron T. Wang, Zeyu Michael Li, Lawrence Carin, and Xiang Cheng. 2025. In-Context Semi-Supervised Learning. *arXiv preprint arXiv:2512.15934* (2025). arXiv:2512.15934 https://arxiv.org/abs/2512.15934

[10] Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. Enhancing Critical Thinking in Education by means of a Socratic Chatbot. *arXiv preprint arXiv:2409.05511* (2024). arXiv:2409.05511 https://arxiv.org/abs/2409.05511

[11] Vanessa Figueiredo. 2025. Fuzzy, Symbolic, and Contextual: Enhancing LLM Instruction via Cognitive Scaffolding. *arXiv preprint arXiv:2508.21204* (2025). arXiv:2508.21204 https://arxiv.org/abs/2508.21204

[12] A.C. Graesser, P. Chipman, B.C. Haynes, and A. Olney. 2005. AutoTutor: An Intelligent Tutoring System With Mixed-Initiative Dialogue. *IEEE Transactions on Education* 48, 4 (Nov 2005), 612–618. https://doi.org/10.1109/te.2005.856149

[13] Liqun He and Jiaqi Xu. 2025. Automated Classification of Tutors' Dialogue Acts Using Generative AI: A Case Study Using the CIMA Corpus. *arXiv preprint arXiv:2509.09125* (2025). arXiv:2509.09125 https://arxiv.org/abs/2509.09125

[14] Sara Hennessy, Elisa Calcagni, Alvin Leung, and Neil Mercer. 2021. An analysis of the forms of teacher-student dialogue that are most productive for learning. *Language and Education* 37, 2 (Aug 2021), 186–211. https://doi.org/10.1080/09500782.2021.1956943

[15] Sara Hennessy, Christine Howe, Neil Mercer, and Maria Vrikki. 2020. Coding classroom dialogue: Methodological considerations for researchers. *Learning, Culture and Social Interaction* 25 (June 2020), 100404. https://doi.org/10.1016/j.lcsi.2020.100404

[16] Sara Hennessy, Neil Mercer, and Paul Warwick. 2011. A Dialogic Inquiry Approach to Working with Teachers in Developing Classroom Dialogue. *Teachers College Record: The Voice of Scholarship in Education* 113, 9 (Sept 2011), 1906–1959. https://doi.org/10.1177/016146811111300902

[17] MiYoung Lee, Hyewon Kim, and Minjeong Kim. 2014. The effects of Socratic questioning on critical thinking in web-based collaborative learning. *Education as Change* 18, 2 (May 2014), 285–302. https://doi.org/10.1080/16823206.2013.849576

[18] Haoxuan Li, Jifan Yu, Xin Cong, Yang Dang, Daniel Zhang-li, Lu Mi, Yisi Zhan, Huiqin Liu, and Zhiyuan Liu. 2025. Which Type of Students can LLMs Act? Investigating Authentic Simulation with Graph-based Human-AI Collaborative System. *arXiv preprint arXiv:2502.11678* (2025). arXiv:2502.11678 https://arxiv.

org/abs/2502.11678

[19] Yaoyuan Liang, Zhao Yang, Yansong Tang, Jiashuo Fan, Ziran Li, Jingang Wang, Philip H.S. Torr, and Shao-Lun Huang. 2023. LUNA: Language as Continuing Anchors for Referring Expression Comprehension. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*. ACM, 5174–5184. https://doi.org/10.1145/3581783.3612584

[20] Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models. In *Advances in Neural Information Processing Systems 37*. 85693–85721. https://doi.org/10.52202/079017-2721

[21] Ying Liu, Can Li, Ting Zhang, Mei Wang, Qiannan Zhu, Jian Li, and Hua Huang. 2025. Discerning Minds or Generic Tutors? Evaluating Instructional Guidance Capabilities in Socratic LLMs. *arXiv preprint arXiv:2508.06583* (2025). arXiv:2508.06583

[22] Xinyi Lu and Xu Wang. 2024. Generative Students: Using LLM-Simulated Student Profiles to Support Question Item Evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)*. ACM, 16–27. https://doi.org/10.1145/3657604.3662031

[23] Wenting Ma, Olusola O. Adesope, John C. Nesbit, and Qing Liu. 2014. Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology* 106, 4 (Nov 2014), 901–918. https://doi.org/10.1037/a0037123

[24] Yiping Ma, Shiyu Hu, Xuchen Li, Yipei Wang, Yuqing Chen, Shiqing Liu, and Kang Hao Cheong. 2024. When LLMs Learn to be Students: The SOEI Framework for Modeling and Evaluating Virtual Student Agents in Educational Interaction. *arXiv preprint arXiv:2410.15701* (2024). arXiv:2410.15701 https://arxiv.org/abs/2410.15701

[25] Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. 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*. 204–221. https://doi.org/10.18653/v1/2025.emnlp-main.11

[26] Neil Mercer, Sara Hennessy, and Paul Warwick. 2019. Dialogue, thinking together and digital technology in the classroom: Some educational implications of a continuing line of inquiry. *International Journal of Educational Research* 97 (2019), 187–199. https://doi.org/10.1016/j.ijer.2017.08.007

[27] Baptiste Moreau-Pernet. 2026. EvalConvoLearn: An Open-Source Framework for Evaluating Grounded Learner Simulations in Tutoring Conversations. *arXiv preprint arXiv:2608.07497* (2026). arXiv:2608.07497 https://arxiv.org/abs/2608.07497

[28] Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. AutoTutor meets Large Language Models: A Language Model Tutor with Rich Pedagogy and Guardrails. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24)*. ACM, 5–15. https://doi.org/10.1145/3657604.3662041

[29] Zachary A. Pardos and Shreya Bhandari. 2023. Learning gain differences between ChatGPT and human tutor generated algebra hints. *arXiv preprint arXiv:2302.06871* (2023). arXiv:2302.06871 https://arxiv.org/abs/2302.06871

[30] Vasile Rus, Dan Stefanescu, William Baggett, Nobal Niraula, Don Franceschetti, and Arthur C. Graesser. 2014. *Macro-adaptation in Conversational Intelligent Tutoring Matters*. Springer International Publishing, 242–247. https://doi.org/10.1007/978-3-319-07221-0_29

[31] Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. 2026. Simulated Students in Tutoring Dialogues: Substance or Illusion?. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 42349–42385. https://doi.org/10.18653/v1/2026.acl-long.1960

[32] Saiying Steenbergen-Hu and Harris Cooper. 2014. A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology* 106, 2 (May 2014), 331–347. https://doi.org/10.1037/a0034752

[33] Albert L. Stevens and Allan Collins. 1977. *The Goal Structure of a Socratic Tutor*. https://doi.org/10.21236/ada038359

[34] Danielle R. Thomas, Conrad Borchers, Kirk P. Vanacore, Kenneth R. Koedinger, and René F. Kizilcec. 2026. Modernizing Ground Truth: Four Shifts Toward Improving Reliability and Validity in AI in Education. *arXiv preprint arXiv:2603.29141* (2026). arXiv:2603.29141 https://arxiv.org/abs/2603.29141

[35] Ahmed Tlili, Soheil Salha, Huanhuan Wang, and Ronghuai Huang. 2023. Intelligent Tutoring Systems Examined in Social Experiments—Is the Magic Gone? A Meta-Analysis. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. IEEE, 50–54. https://doi.org/10.1109/icalt58122.2023.00020

[36] Sifiso Vilakati. 2026. Scaling Socratic Statistics Instruction Through Large Language Models. (Oct 2026). https://doi.org/10.20944/preprints202610.0576.v1

[37] Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2025. Large Language Models for Education: A survey and outlook. *IEEE Signal Processing Magazine* 42, 6 (Nov 2025), 51–63. https://doi.org/10.1109/msp.2025.3594309

[38] Rasil Warnakulasooriya. 2006. Evidence of Problem-Solving Transfer in Web-Based Socratic Tutor. In *AIP Conference Proceedings*, Vol. 818. AIP, 41–44. https://doi.org/10.1063/1.2177018

[39] Rupert Wegerif. 2008. Dialogic or dialectic? The significance of ontological assumptions in research on educational dialogue. *British Educational Research Journal* 34, 3 (June 2008), 347–361. https://doi.org/10.1080/01411920701532228

[40] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2023. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* 55, 1 (Aug 2023), 90–112. https://doi.org/10.1111/bjet.13370

[41] Tingyang Yu, Chanakya Ekbote, Nikita Morozov, Antonio Herrera, Jiashuo Fan, Albert Dominguez Mantes, Salvatore Novello, Hilal Lashuel, Pascal Frossard, Stéphane d'Ascoli, Gioele La Manno, and Maria Brbić. 2025. Tissue Reconstruction with Generative AI. (Feb 2025). https://doi.org/10.1101/2025.02.13.638045

[42] Zhihao Yuan, Yunze Xiao, Ming Li, Weihao Xuan, Richard Tong, Mona Diab, and Tom Mitchell. 2026. Towards Valid Student Simulation with Large Language Models. *arXiv preprint arXiv:2601.05473* (2026). arXiv:2601.05473 https://arxiv.org/abs/2601.05473

[43] Murong Yue, Wenhan Lyu, Jennifer Suh, Yixuan Zhang, and Ziyu Yao. 2024. MathVC: An LLM-Simulated Multi-Character Virtual Classroom for Mathematics Education. *arXiv preprint arXiv:2404.06711* (2024). arXiv:2404.06711 https://arxiv.org/abs/2404.06711