

Can LLMs Interpret Psychometric Parameters? Quantifying "Over-Knowledge" in LLM Student Simulation

Kazunori Fukuhara
Stanford University
kazf28@stanford.edu

Benjamin Domingue
Stanford University
bdomingu@stanford.edu

Abstract

LLM simulation suffers from an “over-knowledge problem”, where models perform too well to represent struggling learners. We compare persona-, IRT-, and CDM-based prompting for student simulation and measure how well methods reflect expected behaviors and quantify this bias. Findings show LLMs follow IRT parameters, yet struggle to simulate real abilities.

1 Introduction

Educational assessments increasingly leverage AI to simulate student responses for validating psychometric models and testing assessment designs without large-scale field trials (Liu et al., 2024). Early studies show LLMs can approximate student performance: GPT-3.5 yielded monotonically increasing accuracy across ability levels (Benedetto et al., 2024), and GPT-4 generated response patterns correlating strongly with real learners (Lu and Wang, 2024). However, LLMs suffer from the “over-knowledge problem”—they tend to produce overly correct responses even when prompted to simulate struggling learners (Li et al., 2025).

To address this, we apply Item Response Theory (IRT) and Cognitive Diagnostic Models (CDMs) as psychometric frameworks alongside baseline persona prompting. IRT models a continuous latent ability governing item performance (Bock and Aitkin, 1981), while CDMs model discrete skill mastery profiles (de la Torre, 2011). Persona prompting serves as a baseline accessible to teachers without psychometric expertise, while IRT- and CDM-backed prompts offer finer control for those with student response data (Hase and Kuhl, 2024). Comparing these three conditions reveals when and how the over-knowledge problem arises.

We generate 250 synthetic student responses based on real MCQ data from IGS’s e-Spire TOEFL reading platform (Japanese middle and

high school students). Two simulation phases examine (1) LLM sensitivity to psychometric parameters and quantification of over-knowledge, and (2) fidelity of simulated responses across prompting conditions—evaluated by correctness alignment, parameter sensitivity, and parameter recovery. This study addresses the following research questions:

- **RQ1:** To what extent can LLMs “interpret” psychometrics parameters (use parameters to generate student response data that resemble the real student response data) and how can we measure the “over-knowledge problem”
- **RQ2:** How do psychometric models (IRT/CDM) control LLM-simulated students compared to baseline persona prompting?

This study makes the following contributions:

- Empirically evaluate LLM sensitivity to numerical IRT parameters vs. natural language persona inputs
- Identify and quantify the “over-knowledge problem” in student simulation settings
- Provide practical guidance for choosing prompting frameworks in LLM-driven assessment design

2 Related Work

2.1 Student Simulation in Pre-LLM Era

Earlier intelligent tutoring systems used rule-based and ML-driven models to generate synthetic student behaviors—for example, SimStudent represents a student as a production rule system that learns from examples and produces realistic error patterns (Matsuda et al., 2023; Srivatsa et al., 2025). These approaches required extensive domain knowledge engineering per scenario, motivating the shift to LLMs for more flexible student simulation.

2.2 Student Profile-Based LLM-Student Simulation

Lu and Wang (2024) introduced Generative Students, using GPT-4 with KLI-based knowledge component profiles (Koedinger et al., 2012) to simulate 45 students answering MCQs; item difficulties derived from simulated responses correlated with real students at $r = 0.72$. However, a challenge is that LLMs are trained to be knowledgeable, producing overly proficient answers even when prompted to simulate struggling students (Wu et al., 2025). To address this, Wu et al. (2025) proposed a training-free pipeline that builds a prototype from past student records and iteratively refines LLM answers to better mirror student mistakes.

2.3 LLMs and Psychometric Models

Several studies treat LLMs as synthetic test-takers to generate response data. Liu et al. (2024) found GPT-3.5/4 responses reproduced item difficulty and discrimination statistics closely matching a 50-student field test, and Zelikman et al. (2023) fine-tuned an LLM to estimate item difficulty via IRT, producing test forms correlating with human-validated versions at $r = 0.93$.

Srivatsa et al. (2025) applied IRT to compare LLMs against NAEP student performance, finding that general LLMs far outperformed grade-level averages; grade persona prompts partially closed the gap but could not precisely match target ability levels—underscoring the need for principled methods to tie LLM outputs to quantitative ability values. Scarlatos et al. (2025) addressed this with SMART, fine-tuning an LLM on student responses conditioned on IRT ability values, then applying Direct Preference Optimization to align generated answers with the intended ability level. This improved fidelity and enabled accurate item difficulty prediction, particularly in low-data settings.

A key challenge is that LLMs tend to treat numerical values as string tokens rather than magnitudes (Sakamoto and Aizawa, 2021; Wallace et al., 2019), limiting their interpretation of IRT parameters. Our study does not rely on specialized embeddings or fine-tuning; instead, we evaluate off-the-shelf LLM prompting across IRT, CDM, and persona-based conditions to provide practical guidance for researchers and teachers.

3 Data

We leverage data from Institution for a Global Society’s (IGS) online TOEFL reading practice platform (e-Spire), used by Japanese middle and high school students for English proficiency exam preparation. Students complete practice reading problem sets with MCQ comprehension questions following each passage. The dataset comprises responses from 7,779 students across 294 passages and 2,176 unique questions collected between May 2017 and April 2025; students attempted 112.7 questions on average with mean correctness of 0.614.

Reading passages are assigned difficulty based on five Eiken examination categories (1, pre-1, 2, pre-2, 3); we supplement this with Rasch model estimates of individual item difficulty. IGS item designers also categorized each question by type, where each category is presented in Appendix A.

We randomly select 250 students from the dataset. Within this sample, each student attempted 108 questions on average, with a mean of 12.5 attempts per question.

For Rasch model estimation via the *mirt* package (Chalmers, 2012) in R, we retain only each student’s first response per item to satisfy IRT local independence (Edwards et al., 2018) and minimize exposure effects (74.6% of students had responded to the same item more than twice). We further restrict to items attempted by at least 20 students and students who responded to at least 10 items—thresholds that balance estimation stability and data retention, and correspond to the minimum conditions for G-DINA convergence. This yields 250 students and 672 items (670 for CDM estimation). Table A1 presents descriptive statistics, and the Q-matrix required for CDM was constructed by the researchers.

We design 250 distinct LLM student profiles for this study using gpt-4.1-nano (OpenAI, 2024), a powerful and cost-effective model, where the ability parameter from the Rasch model is used to assign LLM with different student profiles.

4 Method

We divide the study into two phases: (1) evaluation of LLMs’ ability to “interpret” IRT and CDM parameters using synthetic psychometric inputs, and (2) simulation of real student responses under persona-, IRT-, and CDM-based prompting. Phase 1 serves as a prerequisite benchmark—if LLMs

cannot meaningfully condition on numerical parameters, Phase 2 simulations would be unreliable.

4.1 Phase 1: Evaluation under Synthetic Psychometric Parameters

Phase 1 assesses whether LLMs can interpret IRT or CDM parameters in prompts without fine-tuning, as target users (teachers) may lack resources for such adaptation. Prompts are provided in Appendix C Boxes 1 and 2.

For the IRT condition, we assign ability values (θ_{LLM}) in increments of 0.1 over $[-3, 3]$, and evaluate responses on 15 items spanning three difficulty levels ($b \in \{-1, 0, 1\}$), with five items per difficulty level and five independent runs per ability-item pairing. This design lets us examine sensitivity to both IRT parameters and question-specific variation. For the CDM condition, we test all $2^7 = 128$ skill mastery combinations for the seven skills in the Q-matrix. Both conditions are evaluated at temperatures 0 and 0.7 to assess stochasticity. We assess IRT condition sensitivity through four analyses:

- **Monotonicity:** Kendall’s τ (Kendall, 1945) between ability-difficulty difference and average correctness (p_i), computed over the pooled sample, by difficulty level, and by individual item.
- **Variance:** Response variability within the same ability-difficulty pair ($\widehat{\text{Var}}_{\text{within}(\theta, b)}$) and within the same ability-item pair ($\widehat{\text{Var}}_{\text{within}(\theta, j)}$), distinguishing item-content effects from parameter-driven effects.
- **Parameter Recovery:** Rasch model fitted to LLM-generated responses; Pearson ρ and RMSE between estimated and input ability values (θ_{LLM} vs. θ_{given}). Responses are aggregated across runs and binarized at 0.5 before fitting.
- **Correctness Transition Point:** Novel metric described in Section 4.1.1

For the CDM condition, we examine average correctness and variance by skill mastery level.

4.1.1 Correctness Transition Point Analysis

We introduce the correctness transition point: the ability-difficulty difference ($\theta_i - b_j$) at which simulated average correctness first exceeds $\kappa = 0.5$. Under the Rasch model this threshold occurs at

$\theta - b = 0$, so systematic departures from zero quantify over- or under-knowledge in the simulation.

We define θ^* (transition point prior to difficulty adjustment) via linear interpolation between the two adjacent ability values (θ_{i-1}, θ_i) satisfying $p_{i-1} < \kappa \leq p_i$:

$$\theta^* = \theta_{i-1} + (\theta_i - \theta_{i-1}) \cdot \frac{\kappa - p_{i-1}}{p_i - p_{i-1}}. \quad (1)$$

This yields a continuous estimate invariant to grid spacing. We compute θ^* at both the item level and the difficulty level (averaging across items sharing the same b). The transition point is then $\theta^* - b$, where values above zero indicate over-knowledge. We estimate 95% confidence intervals via trial-level bootstrapping ($B = 1000$ replicates).

4.2 Phase 2: Full Student Simulation with LLMs

We run three prompting conditions—persona, IRT, and CDM—each using few-shot prompting with three prior student responses as examples alongside the passage, question, answer options, and condition-specific profile information.

In all conditions, the LLM is prompted to output a correctness probability for each item; the final response is determined by comparing this probability to a correctness threshold sampled from Beta(7, 3), centering the boundary near 0.7. This calibrates for LLMs’ tendency to overestimate correctness and introduces variability consistent with the model’s internal uncertainty (Figure B1).

4.2.1 Baseline Persona Simulation

Each student is described qualitatively by performance level ("low", "medium", or "high") in each of the nine question categories, based on correctness thresholds of $[0, 0.34)$, $[0.34, 0.67)$, and $[0.67, 1]$ (Appendix C Box 5).

4.2.2 IRT-Based Simulation

For the IRT condition, the LLM is given the student’s Rasch ability and the item’s difficulty and tasked with determining whether the student would answer correctly (Appendix C Box 6).

4.2.3 CDM-Based Simulation

We fit the G-DINA model (Ma and de la Torre, 2020; Li and Suen, 2013) to estimate students’ skill mastery profiles, then adapt the IRT prompting approach by replacing ability/difficulty with mastery

profiles and Q-matrix skill requirements (Appendix C Box 7).

4.2.4 Phase 2 Evaluation Metrics

Correctness and mistake alignment evaluate how well simulated responses reproduce real student correctness rates and error patterns (MCQ option selection when both student and simulation answer incorrectly).

Parameter sensitivity examines whether simulated correctness coherently tracks assigned profiles: under persona prompting, whether high-proficiency students outperform low-proficiency students by category; under IRT, whether correctness increases when ability exceeds difficulty; under CDM, whether correctness rises with the proportion of required skills mastered.

Parameter recovery evaluates whether IRT/CDM models fitted to LLM-generated data recover the original input parameters. For IRT, recovery is assessed via Pearson ρ and RMSE on ability estimates; for CDM, via the proportion of skill mastery statuses correctly classified relative to the real student profiles used in simulation.

5 Results

We present Phase 1 results first, then Phase 2. Temperature settings (0 and 0.7) produced largely consistent results; we report for $temp = 0.7$ unless otherwise noted.

5.1 Phase 1: Evaluation Under Synthetic Psychometric Parameters

5.1.1 IRT Evaluation

Kendall's τ between ability-difficulty difference and average correctness was 0.674 overall, with comparable values across difficulty levels (Easy: 0.717, Medium: 0.676, Hard: 0.686), indicating consistent monotonic sensitivity to IRT parameters.

Within-item variance ($\widehat{\text{Var}}_{\text{within}(\theta,j)}$) and within-difficulty variance ($\widehat{\text{Var}}_{\text{within}(\theta,b)}$) were both near zero across all conditions (Table A2), indicating minimal stochasticity across repeated runs.

Parameter recovery yielded $\rho = 0.937$ (95% CI: 0.897–0.962), indicating strong correlation between input and recovered ability values.

Transition point analysis (Figure B2) shows that the shift from predominantly incorrect to correct responses occurs at ability-difficulty differences of approximately -0.5 for medium and hard items and -0.093 for easy items—all negative,

indicating slight over-knowledge—i.e., responses were correct at lower ability levels than expected—across all item difficulty levels.

5.1.2 CDM Evaluation

The CDM condition revealed a markedly different pattern: the LLM treated skill mastery conjunctively, producing near-perfect correctness only when all required skills were mastered and near-zero correctness otherwise—even when a student mastered all but one skill (Table A3). This rigid all-or-nothing behavior deviates from the expected response given the Q matrix.

5.2 Phase 2: Full Student Simulation

Table A4 reports correctness alignment (whether correctness matches between simulation and real student) across the three conditions using 32,218 real student responses from 250 students. For IRT and CDM, the alignment between real-student response and LLM response from IRT/CDM-based prompts are shown. The overall accuracy remains low around 0.6, where 1 indicates perfect alignment

5.2.1 Correctness and Mistake Alignment

Overall correctness alignment is low across all conditions (0.554–0.598), only modestly above the 0.5 guessing baseline (Table A4). For reference, IRT-predicted correctness aligns at 0.733 and G-DINA at 0.765. The InterModel Vigorish (IMV) (Domingue et al., 2021), which quantifies whether LLM-generated responses predict real student responses better or worse relative to the IRT-based estimates, was -0.022 , indicating LLM-simulated responses offered less predictive value than IRT estimates alone. The baseline condition produced 96.3% simulated correctness against 54.0% in real data, reflecting severe over-knowledge; IRT (66.0%) and CDM (50.2%) were substantially closer to real students.

Mistake alignment was 0.40 for IRT and CDM, and 0.50 for the persona-prompting, modestly above the random-guessing baseline of 0.33 for two-distractor items.

5.2.2 Baseline Simulation

Simulated correctness was largely monotonic across proficiency levels within each question category (Figure B3), with all pairwise differences significant at $\alpha = .05$. Minor violations occurred in *Reference* (medium to high: 0.952 $\rightarrow$ 0.857) and *Simplification* (medium to high: 0.952 $\rightarrow$ 0.910).

5.2.3 IRT Simulation

Kendall's τ between ability-difficulty difference and simulated correctness in Phase 2 was 0.274 ($p < .01$)—lower than Phase 1 but comparable to $\tau = 0.439$ observed in real student data, suggesting the simulation captures the attenuated signal present in actual responses. Table A5 shows a broadly monotonic pattern, but with notable deviations at extremes: at the lowest ability-difficulty bin, the LLM overestimates correctness relative to real students (0.456 vs. 0.151), while at the highest bin it underestimates (0.883 vs. 0.936).

5.2.4 CDM Simulation

Simulated correctness increased monotonically with the number of required skills mastered across all item groups (Figure B4). However, when full mastery was specified, simulated accuracy (0.836–0.940) exceeded real student accuracy (0.753–0.822), consistent with the over-knowledge pattern observed in Phase 1.

Marginal skill mastery effects were inflated in simulation relative to real students for most skills (Figure B5), with the largest gap for *understanding implicit information* (0.394 vs. 0.168). The exception was *Vocabulary Level 2*, where mastery effects were identical (0.356).

5.3 IRT / CDM Parameter Recovery

For IRT, recovered ability correlated with input values at $\rho = 0.913$ ($p < .01$) with RMSE = 0.407, indicating strong rank-order preservation despite the low item-level correctness alignment. The high correlation suggests the simulation maintained relative ability ordering even when absolute correctness was inflated. For CDM, 34.4% of students had skill mastery profiles exactly recovered, with overall skill-level classification accuracy of 0.807.

6 Discussion

6.1 LLM "Interpretation" of IRT Parameters

Phase 1 provides convergent evidence that LLMs can meaningfully interpret IRT parameters. Kendall's $\tau = 0.674$ indicates a moderately strong monotonic relationship between ability-difficulty and correctness (Cohen, 1988); within-difficulty variance was lower than within-item variance, suggesting LLM responses are governed more by IRT parameters than by surface-level phrasing differences. Parameter recovery ($\rho = 0.937$) further

confirms alignment between simulated and psychometrically predicted responses. This finding partly challenges the claim by Sakamoto and Aizawa (2023) that LLMs struggle to interpret numerical inputs.

6.2 "Over-Knowledge Problem" in LLM-Simulated Students

The transition point analysis quantifies the over-knowledge problem precisely. For medium and hard items, the transition from predominantly incorrect to correct responses occurs at $\theta^* - b = -0.5$ rather than 0 as the Rasch model predicts—meaning the LLM answers correctly more than half the time even when ability is 0.5 units below item difficulty. At $\theta - b = -0.5$ the correctness probability is at 0.378 under the Rasch model, where simulation inflated the probability by around 12.2 percentage point. Easy items show a transition point of -0.1 (Rasch correctness of 0.475), close to expectation. This aligns with prior qualitative observations (Wu et al., 2025) and provides a formal, scale-referenced quantification of the effect.

Phase 2 corroborates this at real-student scale: at the most negative ability-difficulty bin ($-4.1, -1.5$], simulated correctness is 0.456 versus 0.151 for real students (gap of 0.305). The gap narrows as ability approaches difficulty, then reverses at the high end where the LLM underestimates performance (0.883 vs. 0.936), consistent with LLMs struggling at distributional extremes (Bender et al., 2021). Importantly, some of this effect may reflect task context: these are ESL reading items calibrated on Japanese students, so items classified as difficult within this dataset likely represent only moderate challenge for an LLM trained predominantly on English text (Guo et al., 2025).

6.3 IRT Parameter Estimation Quality Limits Phase 2 Fidelity

Phase 2 IRT correctness alignment (0.598) was substantially weaker than Phase 1, but the Rasch model alone also achieved only 0.733 alignment with real students—suggesting the bottleneck lies in parameter estimation quality rather than the simulation itself. The IMV was -0.022 , a small difference indicating LLM simulation tracked the IRT benchmark closely. Figure B6 shows both real and simulated person-fit distributions have a slight left skew and neither achieves uniformly ideal fit, likely due to the imbalance between 250 students and 672 items reducing estimation stability.

6.4 Implications of Baseline Simulation

The baseline condition had the lowest correctness alignment (0.554) but the highest mistake alignment (0.496), suggesting that per-category proficiency labels help the model reproduce student-specific error patterns even without psychometric parameters. A consistent limitation is that proficiency labels are interpreted too extremely: students labeled high proficiency often exceeded 0.9 accuracy within that category, well above the 0.67 empirical threshold used to assign the label. This inflation likely reflects LLMs treating qualitative descriptors in an idealized rather than distributional sense, contributing to the systematic overknowledge observed across conditions.

6.5 Implications of CDM Simulation

Phase 1 revealed a rigid conjunctive behavior: correctness was near zero unless all required skills were mastered. This determinism contradicts the compensatory logic of G-DINA (Li et al., 2016), where partial mastery should yield nonzero success probability. In Phase 2, few-shot prompting partially mitigated this—the LLM occasionally answered correctly with incomplete mastery, producing more random behavior closer to real students.

Nonetheless, mastery effects remained inflated. The correctness gap between full mastery and non-mastery was 0.394 in simulation versus 0.168 in real students for *understanding implicit information*—the largest discrepancy. This likely reflects the binary 0/1 mastery encoding: without graded mastery information, the model cannot capture the partial contributions of individual skills, amplifying mastery effects relative to empirical patterns.

7 Conclusion

This study evaluated LLM-based student simulation under baseline persona, IRT-based, and CDM-based prompting across two phases, while identifying and quantifying the "over-knowledge problem."

Phase 1 provided strong evidence that LLMs can meaningfully interpret IRT parameters. Our novel transition point metric quantified the over-knowledge problem as a 0.5-unit shift on the Rasch ability-difficulty scale for medium and hard items—transforming a previously qualitative observation into a transparent, measurable framework. This metric also holds promise as a reward signal for future fine-tuning aimed at reducing over-knowledge. Under CDM, however, the LLM ad-

hered rigidly to binary skill mastery, producing near-deterministic all-or-nothing responses that deviate from G-DINA's probabilistic structure; few-shot prompting partially mitigated this rigidity.

In Phase 2, correctness alignment was modest across all conditions (0.554–0.598), with IRT-based prompting showing the strongest parameter recovery and most stable behavior at non-extreme ability-difficulty ranges. The results suggest that simulation fidelity is constrained less by the LLM's capacity to interpret psychometric parameters and more by the quality of parameter estimation from the underlying data. With stronger model fit and richer student-level information, IRT-based simulation in particular holds meaningful potential as a practical tool for assessment design.

8 Limitation and Future Directions

First, student ability was treated as static throughout each student's full history on e-Spire, ignoring learning and forgetting over what was sometimes multiple years of platform use. A static IRT estimate approximates ability at a point in time but cannot represent longitudinal change; future work should incorporate dynamic ability models.

Second, the dataset lacked demographic or grade-level metadata, making it difficult to assess whether students engaged with appropriately matched items. The Q-matrix was also constructed from question-type labels and Magoosh annotations rather than provided directly, potentially introducing CDM mis-specification. Additionally, the simulation generates each response independently, whereas real student reasoning is shaped by accumulated experience and prior misconceptions; future frameworks should integrate memory mechanisms to capture cumulative learning trajectories.

Third, the LLM's near-deterministic adherence to binary skill mastery indicators is a fundamental limitation of CDM-based prompting. Representing mastery as 0/1 may trigger rule-based reasoning that precludes the graded, compensatory behavior prescribed by G-DINA. Future work should explore probabilistic mastery representations to better align simulated behavior with CDM theory.

Finally, the pipeline focuses on response rather than the reasoning. Teachers are typically more interested in the misconceptions driving incorrect answers than in binary correctness. Extending the framework to generate interpretable reasoning traces would improve pedagogical utility.

References

- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, New York, NY, USA. Association for Computing Machinery.
- Luca Benedetto, Gianluca Aradelli, Andrea Donvito, Alessandro Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. [Using llms to simulate students' responses to exam questions.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368. Association for Computational Linguistics.
- R. Darrell Bock and Murray Aitkin. 1981. [Marginal maximum likelihood estimation of item parameters: Application of an em algorithm.](#) *Psychometrika*, 46(4):443–459.
- R. Philip Chalmers. 2012. [mirt: A multidimensional item response theory package for the r environment.](#) *Journal of Statistical Software*, 48(6):1–29.
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Jimmy de la Torre. 2011. [The generalized dina model framework.](#) *Psychometrika*, 76(2):179–199.
- Benjamin Domingue, Charles Rahal, Jessica Faul, Jeremy Freese, Kacper Kanopka, Alexandros Rigos, and Anshul Tripathi. 2021. [Inter-model vigorish \(imv\): A novel approach for quantifying predictive accuracy when outcomes are binary.](#) <https://osf.io/gu3ap/>. OSF Preprint.
- Michael C. Edwards, Carrie R. Houts, and Li Cai. 2018. [A diagnostic procedure to detect departures from local independence in item response theory models.](#) *Psychological Methods*, 23(1):138–149.
- Yanzhu Guo, Simone Conia, Zelin Zhou, Min Li, Saloni Potdar, and Henry Xiao. 2025. [Do large language models have an english accent? evaluating and improving the naturalness of multilingual llms.](#) In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3823–3838.
- Anna Hase and Peter Kuhl. 2024. [Teachers' use of data from digital learning platforms for instructional design: A systematic review.](#) *Educational Technology Research and Development*, 72:1925–1945.
- Maurice G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251.
- Kenneth R. Koedinger, Albert T. Corbett, and Charles Perfetti. 2012. [The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning.](#) *Cognitive Science*, 36(5):757–798.
- Hairong Li, Chad V. Hunter, and Pui-Wa Lei. 2016. [The selection of cognitive diagnostic models for a reading comprehension test.](#) *Language Testing*, 33(3):391–409.
- Hongli Li and Hoi Suen. 2013. [Constructing and validating a q-matrix for cognitive diagnostic analyses of a reading test.](#) *Educational Assessment*, 18:1–25.
- Ming Li, Han Chen, Yunze Xiao, Jian Chen, Hong Jiao, and Tianyi Zhou. 2025. [Can llms estimate student struggles? human-ai difficulty alignment with proficiency simulation for item difficulty prediction.](#) *arXiv preprint arXiv:2512.18880*.
- Yutong Liu, Suraj Bhandari, and Zachary A. Pardos. 2024. [Leveraging llm-respondents for item evaluation: A psychometric analysis.](#) *Preprint*, arXiv:2407.10899. ArXiv preprint arXiv:2407.10899.
- Xiaoyu Lu and Xinyu Wang. 2024. [Generative students: Using llm-simulated student profiles to support question item evaluation.](#) *Preprint*, arXiv:2405.11591. ArXiv preprint arXiv:2405.11591.
- Wenchao Ma and Jimmy de la Torre. 2020. [GDINA: An R package for cognitive diagnosis modeling.](#) *Journal of Statistical Software*, 93(14):1–26.
- Noboru Matsuda and 1 others. 2023. Simstudent apprentice learner (recent developments). In *Intelligent Tutoring Systems*. Springer.
- OpenAI. 2024. gpt-4.1-nano. <https://openai.com>. Accessed via OpenAI API under Terms of Use: <https://openai.com/policies/terms-of-use/>.
- Taku Sakamoto and Akiko Aizawa. 2021. [Predicting numerals in natural language text using a language model considering the quantitative aspects of numerals.](#) In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 140–150, Online. Association for Computational Linguistics.
- Taku Sakamoto and Akiko Aizawa. 2023. [Predicting numerals in text using nearest neighbor language models.](#) In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4795–4809, Toronto, Canada. Association for Computational Linguistics.
- Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 2025. [SMART: Simulated students aligned with item response theory for question difficulty prediction.](#) *arXiv preprint arXiv:2507.05129*.
- Aditya Srivatsa and 1 others. 2025. Can LLMs reliably simulate real students' abilities in mathematics and reading? *arXiv preprint*.
- Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. [Do nlp models know numbers? probing numeracy in embeddings.](#) *Preprint*, arXiv:1909.07940.

Tao Wu, Jingyuan Chen, Wang Lin, Mengze Li, Yumeng Zhu, Ang Li, Kun Kuang, and Fei Wu. 2025. [Embracing imperfection: Simulating students with diverse cognitive levels using llm-based agents](#). *Preprint*, arXiv:2505.19997.

Eric Zelikman and 1 others. 2023. Generating and evaluating tests for K-12 students with language model simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*.

A Appendix Tables

- **Fact:** Retrieval of directly stated information in the text [289 questions]
- **Inference:** Draw conclusion beyond what’s stated in the text [61 questions]
- **Insert:** Where to insert a given sentence within a passage [1 question]
- **Negation Fact:** Identification of information that is not directly stated in the text [61 questions]
- **Reference:** What given pronouns are referring to in the text [28 questions]
- **Rhetoric:** What is the author’s tone, purpose and rhetorical strategy [44 questions]
- **Simplification:** Simplify parts of texts into simpler idea or sentence [17 questions]
- **Summary:** Summarize the main ideas in the text [76 questions]
- **Vocabulary:** Identify the synonyms / definitions of the word in the text [155 questions]

Table A1: Descriptive statistics of 250 student samples

	Mean (SD)	Min	Max
Question Attempts per Student	128.9 (137.1)	10	672
Attempts per Question	47.9 (33.8)	20	179
Rasch Ability	0 (0.99)	-1.86	2.24
Rasch Difficulty	-0.16 (0.83)	-3.21	2.46

Table A2: IRT within item and difficulty-level variance. Variance remains nearly 0 across all difficulty and trials

Variance	b	Temp 0.7	Temp 0
within(θ, j)	-1	0.044	0.032
within(θ, j)	0	0.041	0.030
within(θ, j)	1	0.033	0.011
within(θ, b)	-1	0.016	0.024
within(θ, b)	0	0.025	0.032
within(θ, b)	1	0.016	0.011

Table A3: CDM correctness by skill mastery proportion. Result exhibits near-zero correctness unless all required skills are fully mastered

Mastery Prop.	Temp 0.7	Temp 0
0	0.003	0
1/4	0.016	0.006
1/3	0.001	0
1/2	0.004	0
2/3	0.003	0
3/4	0.034	0.016
1	1	1

Table A4: Correctness alignment between LLM simulation response under 3 conditions and real-student responses, with correctness alignment between real-student response and IRT/CDM-predicted responses (non-LLM) in parentheses as comparison

Metrics	Baseline	IRT	CDM
Accuracy	0.554	0.598 (0.733)	0.590 (0.765)
Precision	0.548	0.605 (0.746)	0.629 (0.800)
Recall	0.978	0.739 (0.766)	0.585 (0.752)

Table A5: Average correctness for real students and LLM simulations across approximately equal-sized ability-difficulty bins. A monotonically increasing trend in correctness is expected, with minor misalignment observed.

Diff Bin	Sample Size	Mean Diff	Sim Correctness	Student Correctness
(-4.101, -1.5]	3253	-2.012	0.456	0.151
(-1.5, -0.98]	3201	-1.226	0.448	0.235
(-0.98, -0.55]	3242	-0.754	0.469	0.305
(-0.55, -0.17]	3211	-0.357	0.543	0.374
(-0.17, 0.2]	3212	0.019	0.650	0.475
(0.2, 0.6]	3229	0.405	0.710	0.583
(0.6, 1.0]	3245	0.803	0.792	0.691
(1.0, 1.41]	3204	1.202	0.810	0.785
(1.41, 1.96]	3239	1.677	0.841	0.868
(1.96, 5.25]	3181	2.484	0.883	0.936

B Appendix Figures

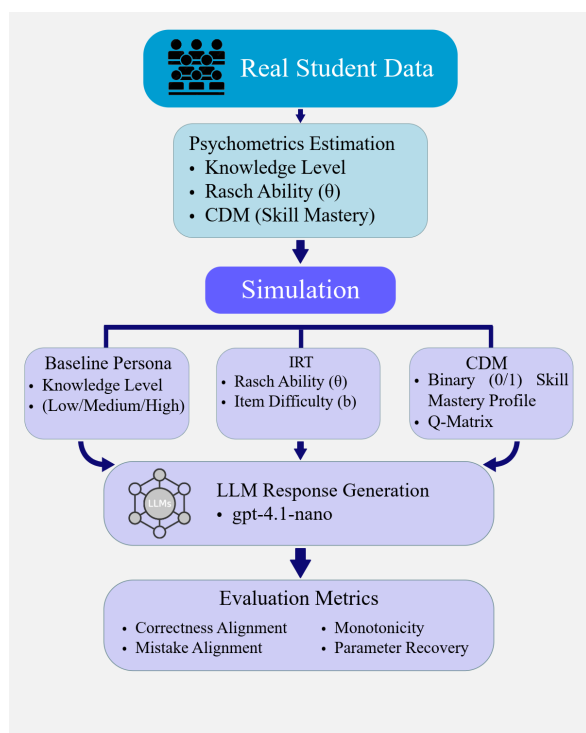


Figure B1: Phase 2 method flow chart

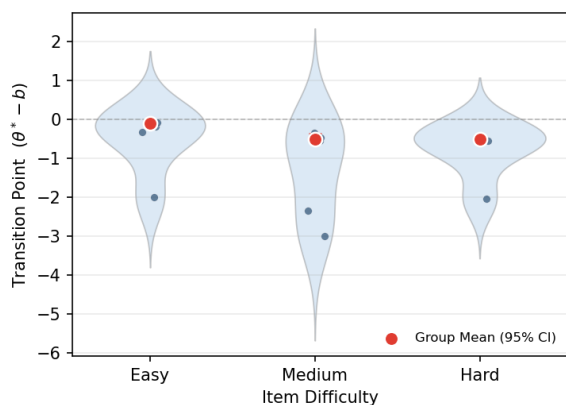


Figure B2: Distribution of transition points across questions within each item difficulty level

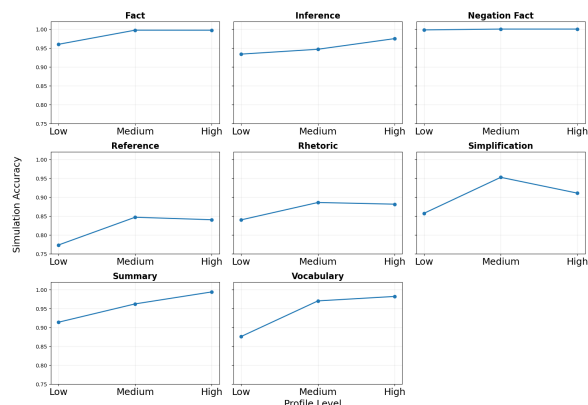


Figure B3: Baseline-condition monotonicity by question kind. Each panel displays simulation accuracy across profile levels.

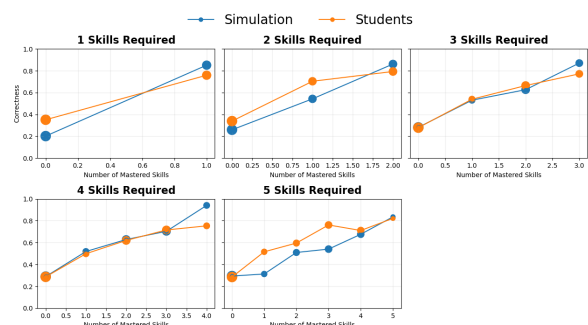


Figure B4: CDM correctness by skill mastery. Each panel corresponds to items requiring 1-5 skills and plots average correctness against the number of mastered skills. Simulation and student performance are shown for comparison, illustrating monotonic patterns within each skill requirement level. Point size reflects the sample size for each skill-mastery level. An upward-sloping relationship is observed across all conditions, consistent with the expected trend.

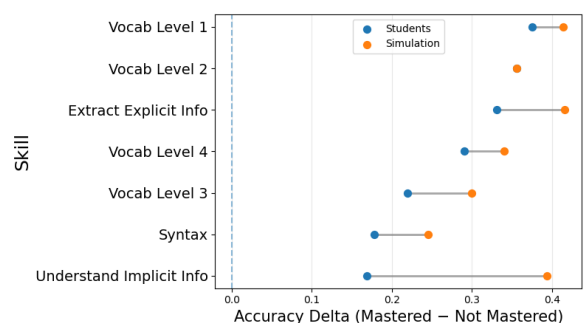


Figure B5: Correctness improvement by CDM skill mastery. We expect small gap between the blue (student response) and the orange dot (simulation response), where in most skills, considerable gap between the two is observed

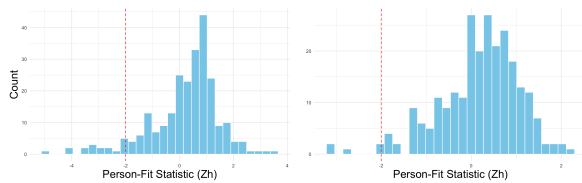


Figure B6: Real student and simulation IRT person-fit statistics distribution. Under high model fit, the two graphs should align, where we observe notable difference between the two

C Example Prompts

This section presents the exact prompts used in the two experimental phases, with question- and student-specific components left as variables.

Box 1: Phase 1 IRT Prompt

Simulate a student taking a
 ↳ multiple-choice reading test.

Student Rasch ability (θ): ABILITY
 Item difficulty (b): DIFFICULTY

Read the passage and answer the question
 ↳ by selecting the option based on the
 ↳ given student ability and difficulty.
 Low ability relative to difficulty means a
 ↳ higher chance of incorrect answers,
 ↳ while high ability means a higher
 ↳ chance of correct answers.
 Choose exactly ONE option. When uncertain
 ↳ about which incorrect option to
 ↳ select, pick randomly among them.

Output format (strict):
 Final answer: <A|B|C|D>

Passage:
 {q['passage']}

Question:
 {q['question']}

Options:
 A) {q['options'][0]}
 B) {q['options'][1]}
 C) {q['options'][2]}
 D) {q['options'][3]}

Box 2: Phase 1 CDM Prompt

Simulate a student taking a
 ↳ multiple-choice reading test under a
 ↳ Cognitive Diagnostic Model (CDM)
 ↳ setting.

Student skill mastery profile (1=mastered,
 ↳ 0=not mastered): SKILL MASTERY

This question requires the following
 ↳ skills (from the Q-matrix): SKILL
 ↳ REQUIREMENT

Guidance:

- If the student has mastered ALL required
 ↳ skills, they are much more likely to
 ↳ answer correctly.
 - If the student is missing ONE OR MORE
 ↳ required skills, they are more likely
 ↳ to answer incorrectly.
 - Outcomes are probabilistic under the
 ↳ G-DINA model: even with all required
 ↳ skills, mistakes can happen; even
 ↳ without having all the required
 ↳ skills, there is some chance of
 ↳ getting an item right.
- Choose exactly ONE option. When uncertain
 ↳ about which incorrect option to
 ↳ select, pick randomly among them.

Output format (strict):
 Final answer: <A|B|C|D>

Passage:
 {q['passage']}

Question:
 {q['question']}

Options:
 A) {q['options'][0]}
 B) {q['options'][1]}
 C) {q['options'][2]}
 D) {q['options'][3]}

Box 3: Phase 2 Common System Prompt

You are simulating a Japanese student
 ↳ learning English on an online
 ↳ platform.

Your PRIMARY goal is to match this
 ↳ student's expected behavior and
 ↳ accuracy level,
 NOT to maximize correctness.

You MUST follow the student's profile and
 ↳ imitate their tendencies.
 You may also receive few-shot history (if
 ↳ provided). Use it to calibrate both
 likelihood of correctness and typical
 ↳ mistakes.

You will be given a value RANDOM_SEED_U01
 ↳ in the user prompt, which is a decimal
 ↳ in [0,1).
 You MUST use it to decide whether the
 ↳ student answers correctly:

Decision rule (MANDATORY):
 1) First decide correctness_probability =
 ↳ P(correct for THIS student on THIS
 ↳ item).
 2) If RANDOM_SEED_U01 <
 ↳ correctness_probability:

output the answer the student would
 ↳ give when correct.
 Else:
 output a plausible wrong option that
 ↳ matches the student's
 ↳ weaknesses (not random).

Important:

- Do NOT choose the correct answer when
 ↳ RANDOM_SEED_U01 >=
- ↳ correctness_probability.
- Do NOT choose an incorrect answer when
 ↳ RANDOM_SEED_U01 <
- ↳ correctness_probability.
- Incorrect answers should reflect
 ↳ realistic misunderstandings.

Output rules:

- Provide brief student-like reasoning
 ↳ (short, imperfect).
 - Provide correctness_probability (0 to
 ↳ 1).
 - Provide exactly ONE final answer in
 ↳ "answer".
 - Respond ONLY in the following JSON
 ↳ format:
- ```
{
 "chain_of_thought": "<BRIEF
 ↳ STUDENT-LIKE RATIONALE>",
 "correctness_probability": <DECIMAL
 ↳ BETWEEN 0 AND 1>,
 "answer": "<ANSWER OPTION>"
}
```

- Inference: low/1/C
- Insert: nan
- Negation Fact: low/1/C
- Reference: high/3/A
- Rhetoric: medium/2/B
- Simplification: nan
- Summary: high/3/A
- Vocabulary: medium/2/B

#### Baseline guidance:

- For the question's category, use the  
 ↳ student's level to infer likelihood of  
 ↳ correctness.
- Use the student's level for question's  
 ↳ category and your judgment on question  
 ↳ difficulty to set  
 ↳ correctness\_probability.
- If the student's category profile  
 ↳ contains many \*\*low/1/C level\*\*,  
 ↳ correctness\_probability should  
 ↳ generally be \*\*below 0.\*\*.
- If the student's category profile  
 ↳ contains many \*\*medium/2/B level\*\*,  
 ↳ correctness\_probability should  
 ↳ generally be \*\*around 0.6\*\*  
 ↳ (uncertain).
- If the student's category profile  
 ↳ contains many \*\*high/3/A level\*\*,  
 ↳ correctness\_probability should  
 ↳ generally be \*\*above 0.7\*\*.

### Box 4: Phase 2 Common User Prompt

RANDOM\_SEED\_U01: CORRECTNESS THRESHOLD

Based on the passage, answer the following  
 ↳ question.

Question: QUESTION

Answer options:

- A) OPTION 1
- B) OPTION 2
- C) OPTION 3
- D) OPTION 4

Respond ONLY in the following JSON format:

```
{
 "chain_of_thought": "<BRIEF
 ↳ STUDENT-LIKE RATIONALE>",
 "correctness_probability": <DECIMAL
 ↳ BETWEEN 0 AND 1>,
 "answer": "<ANSWER OPTION>"
}
```

### Box 6: Phase 2 IRT-Based Prompt

Student profile (IRT):

- English reading ability (theta): ABILITY

IRT guidance (MUST FOLLOW):

- Compare theta and item difficulty (b).
  - If theta is lower than b,  
 ↳ correctness\_probability should  
 ↳ generally be below 0.5, where bigger  
 ↳ the difference, lower the  
 ↳ correctness\_probability.
  - If theta is higher than b,  
 ↳ correctness\_probability should  
 ↳ generally be above 0.5, where bigger  
 ↳ the difference, higher the  
 ↳ correctness\_probability.
  - If close, around 0.5 (uncertain).
- Use few-shot history (if provided) to  
 ↳ calibrate.

Here is the reading passage: PASSAGE

Here Few-shot examples (previous real  
 ↳ student responses): EXAMPLES

Item difficulty (b): DIFFICULTY

### Box 5: Phase 2 Persona-Based Prompt

Student profile (Baseline):

- Estimated performance level by question  
 ↳ category (low/medium/high) (1/2/3)  
 ↳ (C/B/A) - (high, 3 and A represents  
 ↳ the highest level):
- Fact: medium/2/B

### Box 7: Phase 2 CDM-Based Prompt

Student profile (CDM):

- Skill mastery (1 = mastered, 0 = not  
 ↳ mastered):
- Syntax: 0/1
- Extract Explicit Information: 0/1

- Understand Implicit Information: 0/1
- Vocabulary Level 1: 0/1
- Vocabulary Level 2: 0/1
- Vocabulary Level 3: 0/1
- Vocabulary Level 4: 0/1

CDM guidance (G-DINA model):

- Compare the required skills with the
  - ↳ student's mastery profile above.
- If **all** required skills are mastered
  - ↳ (all = 1)\*\*:
    - correctness\_probability should
      - ↳ generally be **above 0.7**.
- If **most** required skills are mastered,
  - ↳ but one is missing\*\*:
    - correctness\_probability should
      - ↳ generally be **around 0.4-0.6**
      - ↳ (uncertain).
- If **multiple** required skills are not
  - ↳ mastered\*\*:
    - correctness\_probability should
      - ↳ generally be **below 0.4**.
- If a **key skill** (central to solving
  - ↳ the item) is not mastered:
    - correctness\_probability should be
      - ↳ **substantially lower**, even if
      - ↳ other skills are mastered.
- Calibrate with few-shot examples

Here is the reading passage: PASSAGE

Required skills for this question:

- SKILL1
- SKILL2