LLM Evaluation on Unseen Questions: Contextual Multidimensional IRT Model

Ergan Shang [1] **Weijing Tang** [1] **Yinqiu He** [2]

Abstract

Evaluation of large language models (LLMs) increasingly requires predicting how a model will perform on new questions or tasks before collecting large amounts of new annotations. This problem is challenging because question difficulty, scenario, and underlying capability demands can vary substantially. Simple retrospective averages may confound model ability with item characteristics. In this paper, we study a model-based evaluation framework that combines multidimensional item response theory model with question contexts to predict LLM performance on unseen questions. The framework represents LLMs through latent capability profiles while using question content to inform item characteristics, allowing information to transfer beyond previously observed items. Empirically, we find that for within-scenario evaluation, incorporating question embeddings improves prediction relative to model-free baselines, and that multidimensional latent structure provides a richer description of capability variation than unidimensional alternatives. At the same time, our results reveal an important limitation that the generalizability does not necessarily translate into reliable prediction under cross-scenario shift. These findings suggest that context-aware psychometric modeling is a promising direction for efficient and interpretable LLM evaluation, while also highlighting cross-scenario generalization as a central open challenge.

1. Introduction

Evaluation of large language models (LLMs) is increasingly shaped by rapid model iteration, evolving benchmarks, and deployment settings that change faster than full-scale annotation pipelines can keep up (Liang et al., 2023; You et al., 2024; White et al., 2025; Shang & Truzzi, 2026). In this setting, a central question is no longer only how a model performed on a fixed benchmark, but how well observed evaluation results can predict behavior on new questions. Reliable prediction on unseen questions is crucial for deployment decisions, because it helps distinguish genuine generalization from benchmark-specific tuning and supports more trustworthy model selection and risk assessment (Truong et al., 2025; Li et al., 2025). This is especially important in high-stakes or rapidly changing applications, where collecting new gold-standard labels may require costly human annotators or expensive judge models (Pacchiardi et al., 2024). In such cases, leveraging existing or small-scale evaluations to predict performance on new questions can substantially improve evaluation efficiency and guide resource allocation (Maia Polo et al., 2024; Wang et al., 2025).

Reliable prediction is challenging, however, because benchmark items are heterogeneous. Questions can vary substantially in difficulty, context, and the underlying capabilities they demand (Zhuang et al., 2025). As a result, simply averaging correctness across questions can confound model capabilities with question characteristics, obscuring structures that matter for extrapolation to unseen questions (Lalor et al., 2016).

Item response theory (IRT) serves as a natural approach to address this problem (Baker, 2001; Chen et al., 2025). Prior studies have used IRT to build evaluation scales, estimate latent item parameters from model response patterns, analyze benchmark quality, and enable more reliable or amortized evaluation (Lalor et al., 2016; 2019; 2024; Maia Polo et al., 2024). More recently, several works have combined IRT modeling with question embeddings to estimate item characteristics and support prediction on previously unseen questions (McCarthy et al., 2021; Marinho et al., 2023; Truong et al., 2025; Tang et al., 2026). However, most existing approaches remain largely based on unidimensional ability, which may not capture the multiple latent capabilities that drive LLM performance. They also leave open a harder question: *whether predictive evaluation can generalize across heterogeneous scenarios* rather than only to held-out questions within the same benchmark setting. This broader scope of generalization is especially important when evaluating LLMs on new tasks or benchmarks (Phan et al., 2025; Li et al., 2026; Zhou et al., 2026).

[1]Department of Statistics and Data Science, Carnegie Mellon University [2]Department of Statistics, University of Madison. Correspondence to: Weijing Tang <weijingt@andrew.cmu.edu>, Yinqiu He <yinqiu.he@wisc.edu>.

Technical report. August 25, 2026.

In this work, we study LLM evaluation on unseen questions through a contextual multidimensional IRT (MIRT) model. Our model jointly captures question characteristics and model capabilities, allowing semantic information from question text to inform item parameters while providing a richer description of latent capability structure.

Our main findings are threefold. First, incorporating question embeddings into MIRT models improves predictive performance relative to model-free alternatives. Second, allowing multiple latent dimensions yields a richer description of model capability than a purely unidimensional approach. Third, although the proposed framework performs encouragingly when predicting held-out questions within the same scenario, its cross-scenario performance has larger variation, suggesting that transfer learning across domains remains a major open challenge. In summary, these results show context-aware psychometric modeling is a promising direction for efficient and interpretable LLM evaluation, while cautioning that within-scenario predictive success may not necessarily translate to robust cross-scenario generalization.

1.1. Related work

Benchmark-based evaluation remains dominant for comparing LLMs, but its limitations in both generalizability and efficiency are increasingly recognized. Prior studies have argued that conclusions drawn from static benchmarks may not directly transfer to changing evaluation protocols (Alzahrani et al., 2024) or scenarios (Kiela et al., 2021; Lin et al., 2025). Meanwhile, a growing line of work seeks to reduce evaluation cost by recovering full-benchmark conclusions from only a small subset of instances (Maia Polo et al., 2024; Pacchiardi et al., 2024; Li et al., 2025; Wang et al., 2025; Zhong et al., 2025). Our work shares the goal of moving beyond retrospective scoring on a fixed benchmark, but studies a structured model-based approach for predicting performance on unseen questions.

Achieving this goal requires separating model-side abilities from heterogeneous question-side characteristics. IRT has long offered a principled way to disentangle item properties from participant abilities in psychometrics (Baker, 2001; Cai et al., 2016). In natural language processing (NLP), Lalor et al. (2016; 2019) introduce IRT-based evaluation scales. The recent tutorial by Lalor et al. (2024) highlights growing interest in IRT as a general framework for model assessment in language technology. In the LLM setting, an emerging number of studies have used IRT-based modeling to study benchmark measurement quality or interpretations (Zhou et al., 2025; Yao et al., 2025; Cai et al., 2025).

A parallel line of work leverages question texts or embedding-based features to model item characteristics and support generalization to unseen items. In educational assessment, such approaches have been used for predicting difficulty (AlKhuzaey et al., 2024; Marinho et al., 2023) or unseen items (McCarthy et al., 2021; Khan et al., 2025). For LLM evaluation, the work most closely related to ours is Truong et al. (2025), who combine Rasch-style psychometric modeling and question embeddings to support reliable and efficient amortized evaluation.

Our method builds on this line of work but differs in two key respects. First, we explicitly model multivariate latent capability dimensions of LLMs rather than relying on a unidimensional Rasch-model view or hard-to-interpret blackbox predictors. Second, we focus on the generalizability of evaluation prediction on unseen questions, including the more difficult setting of cross-scenario transfer rather than only efficient estimation on a fixed benchmark.

2. Method

To predict evaluations of LLMs on unseen questions, we use the following contextual multidimensional item response theory (C-MIRT) model that separates model-side latent abilities from question-side characteristics across scenarios.

C-MIRT model. Consider a benchmark dataset where models are evaluated across S different scenarios, i.e., different task types or contextual domains. For each scenario $s \in [S]$, suppose n LLMs are evaluated on p_s questions. Let $y_{ij}^{(s)} \in \{0,1\}$ denote whether model i answers question j correctly in scenario s. Each question j is associated with a contextual embedding $e_j \in \mathbb{R}^d$. To model performance on these questions, we map this embedding into an r-dimensional latent question representation through a feature map $\phi^{(s)} : \mathbb{R}^d \to \mathbb{R}^r$. Then, the response probability in the C-MIRT is modeled through a logistic link as

$$y_{ij}^{(s)} \mid \theta_{ij}^{(s)} \sim \text{Bernoulli}\Big(\sigma(\theta_{ij}^{(s)})\Big), \qquad \sigma(x) = \frac{1}{1+e^{-x}},$$

with the parameter

$$\theta_{ij}^{(s)} = \alpha_i^{(s)} + {u_i^{(s)}}^\top \phi^{(s)}(e_j). \tag{1}$$

Here $\alpha_i^{(s)} \in \mathbb{R}$ is a scenario-specific intercept for model i, capturing its baseline success rate in scenario s, and $u_i^{(s)} \in \mathbb{R}^r$ is a scenario-specific latent capability vector. The transformed embedding $\phi^{(s)}(e_j)$ represents the latent characteristics of question j in scenario s. The bilinear form $u_i^{(s)\top}\phi^{(s)}(e_j)$ captures how well the capabilities of model i align with the requirements of question j. Different specifications of $\phi^{(s)}$ can be used to incorporate contextual information about the questions. For example, Tang et al. (2026) models $\phi^{(s)}$ as a smooth function in a reproducing kernel Hilbert space, so that questions with similar contextual embeddings are encouraged to have similar latent representations.

The C-MIRT formulation goes beyond a unidimensional difficulty-based view by allowing model performance to vary along multiple latent traits. In particular, the C-MIRT model is closely related to the Rasch-style formulation in Truong et al. (2025), which models

$$\theta_{ij}^{(s)} = \alpha_i^{(s)} - \beta_j^{(s)},$$

with a scalar difficulty parameter $\beta_j^{(s)}$ for question j in scenario s. In that formulation, all item heterogeneity is absorbed into a single difficulty dimension. In contrast, C-MIRT in (1) takes a multivariate bilinear form, which allows question characteristics and model capabilities to interact through multiple latent dimensions.

An additional advantage of C-MIRT is that it naturally supports prediction on unseen questions. Once the feature map $\phi^{(s)}$ is learned from training questions in scenario s, a new question can be embedded into the same latent space and used to predict response probabilities. This allows us to investigate the generalization within and across scenarios.

Model estimation. Given a source scenario s, we estimate the model parameters $(\alpha_i^{(s)}, u_i^{(s)}, \phi^{(s)})$ by minimizing the logistic negative log-likelihood under the C-MIRT model, where the feature map $\phi^{(s)}$ is parameterized by a multilayer perceptron. Because the parameters in MIRT are only identifiable up to a linear transformation, we adopt a two-step estimation procedure; details are deferred to the Appendix.

Predicting responses to unseen questions. Suppose $\widehat{\alpha}_i^{(s)}, \widehat{u}_i^{(s)}, \widehat{\phi}^{(s)}$ are estimated from the training split of the source scenario s. For a question j from target scenario t, with embedding e_j, we compute

$$\widehat{\theta}_{ij}^{(s\to t)} = \widehat{\alpha}_i^{(s)} + \widehat{u}_i^{(s)\top} \widehat{\phi}^{(s)}(e_j), \qquad \widehat{p}_{ij}^{(s\to t)} = \sigma\left(\widehat{\theta}_{ij}^{(s\to t)}\right).$$

When $t = s$, this gives within-scenario predictions for previously unseen questions, enabled by the learned feature map $\widehat{\phi}^{(s)}$. When $t \neq s$, this gives cross-scenario predictions. This directly evaluates out-of-scenario generalization, and the performance depends on how well the latent structure learned under scenario s transfers to scenario t, as we demonstrate empirically in Section 3.1.

3. Experiments

We adopt the benchmark setup in Truong et al. (2025), which leveraged 22 datasets from 5 HELM repositories: Classic, Lite, AIR-Bench, Thai Exam, and MMLU. We filtered out scenarios where the question descriptions are incomplete, e.g., questions with missing multiple-choice options. After filtering, we retain 11 scenarios for evaluation. Examples of scenarios include "wikifact" and "math". The number of questions per scenario ranges from 436 to 29407.

We construct a contextual embedding e_j for each benchmarking question using BERT-based language models (Wang et al. (2020); Reimers & Gurevych (2019)).

All methods are evaluated using 5-fold cross-validation. Within each scenario, questions are split into five folds. In each run, one fold, containing 20% of questions, is used for testing, and the remaining four folds, containing 80% of questions, are used for training.

For each training-test scenario pair and each fold, the fitted model produces a predicted score matrix, where rows correspond to n LLMs and columns correspond to $p_{s,test}$ testing questions in scenario s. We evaluate this matrix in two ways. The first is **question-wise AUC**, or column-wise AUC, where for each fixed test question, we compute the AUC using the predicted scores across n LLMs. This reflects how well the predicted scores rank LLMs on the same question. Repeating this over $p_{s,test}$ questions yields a distribution of question-wise AUC values. The second is **LLM-wise AUC**, or row-wise AUC, where for a fixed LLM, we compute the AUC using the predicted scores across $p_{s,test}$ questions. This measures how well the predicted scores rank questions for that LLM. Repeating this over n LLMs yields a distribution of LLM-wise AUC values. To summarize either distribution, we report its 90th percentile.

3.1. Within- and cross-scenario prediction by C-MIRT

This section evaluates the prediction performance of the C-MIRT. Figures 1 and 2 report the mean 90th percentile AUC over the five cross-validation folds. In both figures, rows indicate the training scenario and columns indicate the test scenario. Diagonal entries therefore represent within-scenario prediction, whereas off-diagonal entries represent cross-scenario prediction.

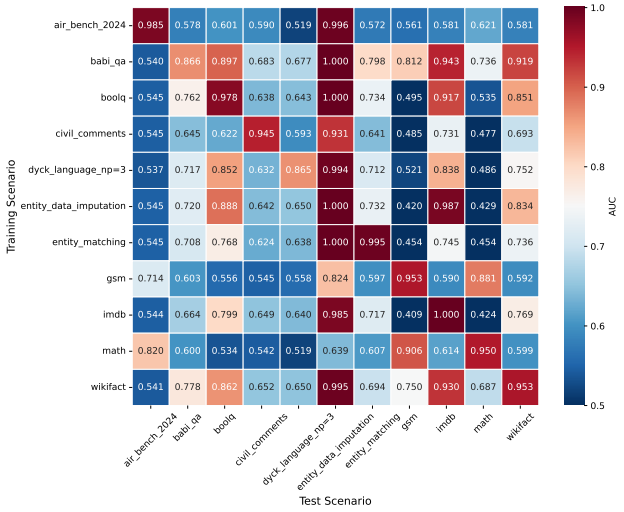

Figure 1. Heatmap of the 90th percentile of question-wise AUC distribution predicted by the C-MIRT method across scenarios.

Figure 1 shows the results for question-wise AUC. The diagonal entries are consistently high, indicating strong within-scenario performance in ranking LLMs for fixed questions. Several off-diagonal entries are also relatively large, suggesting that the learned contextual structure can transfer across scenarios to some extent. For example, training scenarios such as *babi_qa* and *wikifact*, achieve comparatively strong performance on multiple test scenarios. Meanwhile, the substantial variation among off-diagonal cells indicates that cross-scenario transfer depends on the similarity between the source and target scenarios.

Figure 2 shows the results for LLM-wise AUC. Compared with Figure 1, the diagonal entries remain strong, although the off-diagonal entries are much closer to 0.5. This suggests that ranking question difficulty for a fixed LLM is more scenario-specific and less transferable across scenarios than ranking LLMs for a fixed question. Overall, C-MIRT model performs well for within-scenario prediction in both settings, while cross-scenario generalization is notably stronger for question-wise AUC than for LLM-wise AUC.

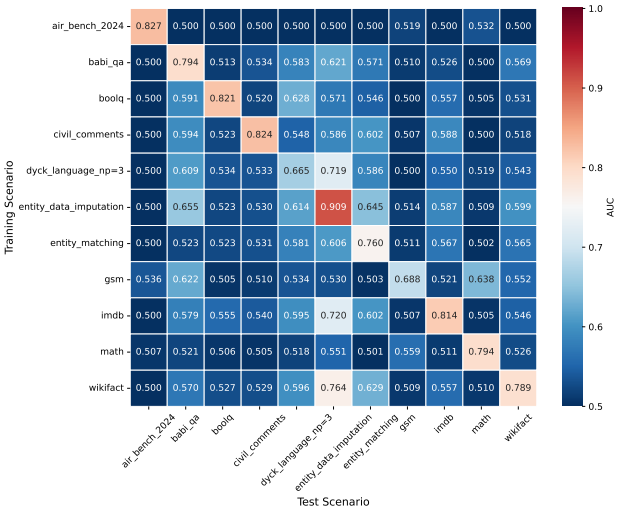

Figure 2. Heatmap of the 90th percentile of LLM-wise AUC distribution predicted by the C-MIRT method across scenarios.

3.2. Comparison with competing methods

We compare C-MIRT with three baselines. The first is the contextual Rasch-style model in Truong et al. (2025). The second is a lasso-penalized logistic regression model with parameters $\theta_{ij}^{(s)} = \gamma_{i0}^{(s)} + \gamma_i^{(s)\top} e_j$. This baseline also uses contextual embeddings e_j, but unlike C-MIRT, does not impose a low-rank latent structure in the coefficient matrix. The third is a simple mean baseline: within each scenario, the predicted score for every test question is set to the average training-set accuracy. Because this baseline assigns the same score to all questions for a given LLM, we use it only for question-wise AUC.

For each method, we compute the 90th percentile AUC in the same way as in Section 3.1. For each pair of methods, 5-fold cross-validation produces 5 paired differences between C-MIRT and the comparator. We assess the significance of these differences using paired t-tests.

In Figures 3 and 4, each heatmap cell corresponds to a scenario–method pair. The number in each cell is the average, across the five folds, of the difference in the 90th percentile of the AUC distribution between C-MIRT and the competing method. A cell is shaded gray when its paired t-test yields a p-value > 0.05, indicating the difference is not statistically significant at the 5% level. Among the remaining cells, red indicates that C-MIRT performs better, whereas blue indicates that the competing method performs better.

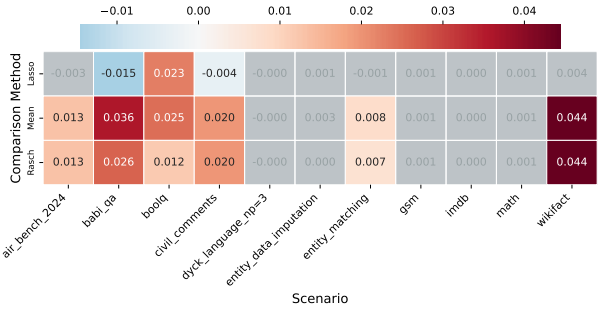

Figure 3. Heatmap of mean differences in the 90th percentile of question-wise AUC between C-MIRT and competing methods across scenarios.

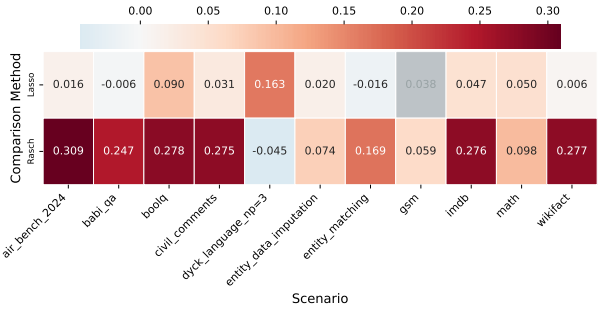

Figure 4. Heatmap of mean differences in the 90th percentile of LLM-wise AUC between C-MIRT and competing methods across scenarios.

Figure 3 shows that for question-wise AUC, C-MIRT consistently outperforms the Mean and Rasch baselines across different scenarios. In contrast, its comparison with the lasso method is more scenario-dependent. Figure 4 shows that, however, for LLM-wise AUC, C-MIRT remains competitive across the majority of scenarios. These results highlight the robustness of the proposed method.

4. Conclusions and future work

This paper proposes to use C-MIRT for predicting evaluations of LLMs on unseen questions. We show that C-MIRT enhances within-scenario prediction, whereas cross-scenario generalizability depends on the source-target scenarios and criterion. It opens an interesting future direction to further investigate new-scenario prediction for LLMs.

Impact Statement

This paper studies a more efficient and interpretable approach to evaluating LLMs on unseen questions. A potential positive impact of this work is that it can reduce the cost of AI evaluation, both in annotation effort and computation, while still providing structured information about model strengths and weaknesses. However, the method also presents risks. Predicted performance may be mistaken for direct evidence of reliability, especially in high-stakes applications, and embedding-based item models may inherit biases from benchmark data or pretrained text representations. Our results also show that cross-scenario prediction is substantially harder than within-scenario prediction, so these methods should not be treated as a substitute for direct testing on representative data. We therefore view this work as a tool for supplementing evaluation pipelines, not replacing careful benchmark design, independent validation, or human oversight.

References

AlKhuzaey, S., Grasso, F., Payne, T. R., and Tamma, V. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914, 2024.

Alzahrani, N., Alyahya, H. A., Alnumay, Y., Alrashed, S., Alsubaie, S., Almushaykeh, Y., Mirza, F., Alotaibi, N., Altwairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024.

Baker, F. B. *The basics of item response theory*. ERIC, 2001.

Cai, L., Choi, K., Hansen, M., and Harrell, L. Item response theory. *Annual Review of Statistics and Its Application*, 3 (1):297–321, 2016.

Cai, P., Cui, C., Polo, F. M., Somerstep, S., Choshen, L., Yurochkin, M., Banerjee, M., Sun, Y., Tan, K. M., and Xu, G. A latent variable framework for scaling laws in large language models. *arXiv preprint arXiv:2512.06553*, 2025.

Chen, Y., Li, X., Liu, J., and Ying, Z. Item response theory—a statistical framework for educational and psychological measurement. *Statistical Science*, 40(2):167–194, 2025.

Khan, A., Li, N., Shen, T., and Rafferty, A. N. Just read the question: Enabling generalization to new assessment items with text awareness. *arXiv preprint arXiv:2507.08154*, 2025.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021.

Lalor, J. P., Wu, H., and Yu, H. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 648–657, 2016.

Lalor, J. P., Wu, H., and Yu, H. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4249–4259, 2019.

Lalor, J. P., Rodriguez, P., Sedoc, J., and Hernández-Orallo, J. Item response theory for natural language processing. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 9–13, 2024.

Li, Y., Ma, J., Ballesteros, M., Benajiba, Y., and Horwood, G. Active evaluation acquisition for efficient LLM benchmarking. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 35581–35602, 2025.

Li, Z., Li, Z., Shi, Y., Wang, R., Yang, J., Liu, Z., Wu, X., Li, A., Yu, Y., Liu, N., et al. Long-horizon-terminal-bench: Testing the limits of agents on long-horizon terminal tasks with dense reward-based grading. *arXiv preprint arXiv:2607.08964*, 2026.

Liang, P., Bommasani, R., Lee, T., et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Lin, B. Y., Deng, Y., Chandu, K., Ravichander, A., Pyatkin, V., Dziri, N., Bras, R. L., and Choi, Y. WildBench: Benchmarking LLMs with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MKEHCx25xp.

Maia Polo, F., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 34303–34326, 2024.

Marinho, W., Clua, E. W. G., Martí, L., and Marinho, K. Predicting item response theory parameters using question statements texts. In *Proceedings of the 13th International Learning Analytics and Knowledge Conference*, pp. 1–10, 2023.

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., and Settles, B. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 883–899, 2021.

Pacchiardi, L., Cheke, L. G., and Hernández-Orallo, J. 100 instances is all you need: predicting the success of a new LLM on unseen data by testing on a few instances. *arXiv preprint arXiv:2409.03563*, 2024.

Phan, L., Gatti, A., Han, Z., Li, N., Hu, J., Zhang, H., Zhang, C. B. C., Shaaban, M., Ling, J., Shi, S., et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Shang, E. and Truzzi, F. S. Erase: Early backpropagation schedule for faster training of modern recommendation systems, 2026. URL https://arxiv.org/abs/2608.18469.

Tang, W., Yuan, M., Xia, Z., and Cai, T. Knowledge-embedded latent projection for robust representation learning. *arXiv preprint arXiv:2602.16709*, 2026.

Truong, S. T., Tu, Y., Liang, P., Li, B., and Koyejo, S. Reliable and efficient amortized model-based evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 60238–60265, 2025.

Wang, S., Wang, C., Fu, W., Min, Y., Feng, M., Guan, I., Hu, X., He, C., Wang, C., Yang, K., Ren, X., Huang, F., Liu, D., and Zhang, L. Rethinking LLM evaluation: Can we evaluate LLMs with 200x less data? *arXiv preprint arXiv:2510.10457*, 2025.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33: 5776–5788, 2020.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Dey, S., Shubh-Agrawal, Sandha, S. S., Naidu, S. V., Hegde, C., LeCun, Y., Goldstein, T., Neiswanger, W., and Goldblum, M. LiveBench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sKYHBTAxVa.

Yao, L. H., Jarvis, N., Zhan, T., Ghosh, S., Liu, L., and Jiang, T. JE-IRT: A geometric lens on LLM abilities through joint embedding item response theory. *arXiv preprint arXiv:2509.22888*, 2025.

You, J., Liu, M., Prabhumoye, S., Patwary, M., Shoeybi, M., and Catanzaro, B. LLM-Evolve: Evaluation for LLM's evolving capability on benchmarks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16937–16942, 2024.

Zhong, X.-X., Yi, C., and Ye, H.-J. Efficient evaluation of large language models via collaborative filtering. *arXiv preprint arXiv:2504.08781*, 2025.

Zhou, H., Huang, H., Zhao, Z., Han, L., Wang, H., Chen, K., Yang, M., Bao, W., Dong, J., Xu, B., Zhu, C., Cao, H., and Zhao, T. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *arXiv preprint arXiv:2505.15055*, 2025.

Zhou, J., Sun, Z., Li, B., Zhou, J., Pan, Y., Wang, H., Ren, H., Jia, X., Zhou, X., Cao, X., Chen, Y., Feng, Y., Wu, J., Zhang, C., Chen, S., Xue, H., You, C., Wang, H., Wu, K., Gao, P., Wu, J., Li, W., Shang, E., Zheng, Q., Zhou, J., Jia, R., Xu, Y., Zhang, H., Ma, X.-H., Cheng, Z., Hao, Y., Mai, L., Ji, X., Zhang, W., Chen, Z., Huang, Y., Wang, C., Hua, W., Hao, Y., Zhai, Y., Zhao, Z., and Xie, J. Asi-bench: At the dawn of artificial superintelligence, 2026. URL https://arxiv.org/abs/2608.17271.

Zhuang, Y., Liu, Q., Pardos, Z., Kyllonen, P. C., Zu, J., Huang, Z., Wang, S., and Chen, E. Position: AI evaluation should learn from how we test humans. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 82483–82508, 2025.

A. Appendix

Matrix representation and model identifiability. For a fixed scenario s, let $Y^{(s)} \in \{0,1\}^{n \times p_s}$ be the response matrix, let $\boldsymbol{\alpha}^{(s)} = (\alpha_1^{(s)}, \ldots, \alpha_n^{(s)})^\top \in \mathbb{R}^n$, let $\mathbf{U}^{(s)} \in \mathbb{R}^{n \times r}$ collect the row vectors ${u_i^{(s)}}^\top$, and let $\mathbf{V}^{(s)} \in \mathbb{R}^{p_s \times r}$ collect the transformed embeddings $\phi^{(s)}(e_j)^\top$.

When the latent question factors are estimated freely rather than parameterized through embeddings, the model reduces to

$$\boldsymbol{\Theta}^{(s)} = \boldsymbol{\alpha}^{(s)} \mathbf{1}_{p_s}^\top + \mathbf{U}^{(s)} \mathbf{V}^{(s)\top}.$$

In that case, to remove the usual rotational ambiguity, one may impose

$$\mathbf{V}^{(s)\top} \mathbf{1}_{p_s} = \mathbf{0}_r, \qquad \mathbf{U}^{(s)\top} \mathbf{U}^{(s)} = \mathbf{V}^{(s)\top} \mathbf{V}^{(s)},$$

under which the factorization is identifiable up to an orthogonal transformation (Tang et al., 2026). Specifically, if there exists another set of parameters $\{\bar{\boldsymbol{\alpha}}^{(s)}, \bar{\mathbf{U}}^{(s)}, \bar{\mathbf{V}}^{(s)}\}$ satisfying the same constraints such that $\boldsymbol{\alpha}^{(s)} \mathbf{1}_{p_s}^\top + \mathbf{U}^{(s)} \mathbf{V}^{(s)\top} = \bar{\boldsymbol{\alpha}} \mathbf{1}_{p_s}^\top + \bar{\mathbf{U}}^{(s)} \bar{\mathbf{V}}^{(s)\top}$, then

$$\boldsymbol{\alpha}^{(s)} = \bar{\boldsymbol{\alpha}}^{(s)}, \quad \mathbf{U}^{(s)} = \bar{\mathbf{U}}^{(s)} \mathbf{O}, \quad \text{and} \quad \mathbf{V}^{(s)} = \bar{\mathbf{V}}^{(s)} \mathbf{O}, \quad \text{for some } \mathbf{O} \in \mathcal{O}(r),$$

where $\mathcal{O}(r)$ collects all orthonormal matrices in $\mathbb{R}^{r \times r}$.

Estimation procedure. Under the Bernoulli model with logistic link, we estimate the C-MIRT parameters using a two-stage procedure. In the first stage, for each scenario s, we fit a low-rank logistic factorization model to obtain latent row and column representations $(\boldsymbol{\alpha}^{(s)}, \mathbf{U}^{(s)}, \mathbf{V}^{(s)})$. In the second stage, we learn the mapping $\phi^{(s)}$ from contextual embeddings to the estimated question factors by supervised regression.

For a fixed scenario s, the negative log-likelihood under the Bernoulli model with logistic link is given by

$$\mathcal{L}^{(s)}\big(\boldsymbol{\alpha}^{(s)}, \mathbf{U}^{(s)}, \mathbf{V}^{(s)}\big) = \sum_{i=1}^{n} \sum_{j \in \mathcal{T}_s} \left[\log\Big(1 + \exp\big(\theta_{ij}^{(s)}\big)\Big) - y_{ij}^{(s)} \theta_{ij}^{(s)} \right], \tag{2}$$

where $\mathcal{T}_s$, the index set, indicates those questions in the training data of scenario s.

To address the identifiability issue, we impose the centering constraint $\mathbf{V}^{(s)\top} \mathbf{1}_{p_s} = \mathbf{0}_r$, and we encourage balanced factorizations through the regularizer

$$g\big(\mathbf{U}^{(s)}, \mathbf{V}^{(s)}\big) = \Big\| \mathbf{U}^{(s)\top} \mathbf{U}^{(s)} - \mathbf{V}^{(s)\top} \mathbf{V}^{(s)} \Big\|_F^2.$$

Accordingly, in stage 1 we solve the regularized optimization problem

$$\min_{\boldsymbol{\alpha}^{(s)}, \mathbf{U}^{(s)}, \mathbf{V}^{(s)}} \quad \mathcal{L}_R^{(s)}\big(\boldsymbol{\alpha}^{(s)}, \mathbf{U}^{(s)}, \mathbf{V}^{(s)}\big) := \mathcal{L}^{(s)}\big(\boldsymbol{\alpha}^{(s)}, \mathbf{U}^{(s)}, \mathbf{V}^{(s)}\big) + \frac{1}{4} g\big(\mathbf{U}^{(s)}, \mathbf{V}^{(s)}\big), \tag{3}$$

subject to

$$\mathbf{V}^{(s)\top} \mathbf{1}_{p_s} = \mathbf{0}_r.$$

We optimize (3) using projected gradient descent. At each iteration, we update $\boldsymbol{\alpha}^{(s)}$, $\mathbf{U}^{(s)}$, and $\mathbf{V}^{(s)}$ by gradient descent, followed by a projection step that enforces $\mathbf{V}^{(s)\top} \mathbf{1}_{p_s} = \mathbf{0}_r$.

After obtaining the stage-1 estimator

$$\widehat{\mathbf{V}}^{(s)} = (\hat{v}_1^\top, \cdots, \hat{v}_j^\top, \cdots)_{j \in \mathcal{T}_s} \in \mathbb{R}^{p_s \times r},$$

we proceed to stage 2 and estimate the feature map $\phi^{(s)}$ by regressing $\hat{v}_j^{(s)}$ on the corresponding contextual embedding e_j. Specifically, we parameterize $\phi^{(s)}$ by a multilayer perceptron and minimize the regression loss

$$\mathcal{L}_{\text{reg}}^{(s)} = \frac{1}{|\mathcal{T}_s^b|} \sum_{j \in \mathcal{T}_s^b} \Big\| \phi^{(s)}(e_j) - \hat{v}_j^{(s)} \Big\|_2^2$$

over mini-batches $\mathcal{T}_s^b \subseteq \mathcal{T}_s$. This yields an estimator $\widehat{\phi}^{(s)}$, which can then be used to predict latent representations for unseen questions and hence out-of-sample response probabilities through

$$\hat{\theta}_{ij}^{(s)} = \hat{\alpha}_i^{(s)} + \hat{u}_i^{(s)\top} \widehat{\phi}^{(s)}(e_j).$$