From Benchmarking to Failure Diagnosis: Diagnosing Classical Chinese Understanding in LLMs via IRT and Error Attribution

Jiong Wang[1,†] **Tengju Li**[1,†] **Shengquan Yu**[2,*]
[1]School of Artificial Intelligence, Beijing Normal University, China
[2]Advanced Innovation Center for Future Education, Beijing Normal University, China
jjoinw@mail.bnu.edu.cn ltj@mail.bnu.edu.cn yusq@bnu.edu.cn

Abstract

Literary Chinese reading comprehension is a challenging task for large language models (LLMs), requiring understanding of implicit meanings, contextual reasoning, and discourse-level integration in a low-resource setting. However, existing evaluations rely primarily on aggregate accuracy, which provides limited insight into model capabilities, error sources, and reliability under uncertainty. We propose **DELit**, a diagnostic evaluation framework that moves beyond aggregate metrics by introducing a hierarchical capability–error taxonomy, a Deep-MIRT-based ability estimation model, and a capability-decoupled residual analysis for behavioral diagnosis. Together, these components enable fine-grained interpretation of model performance at both capability and error-type levels. Experiments on mainstream LLMs reveal that (i) discourse-level understanding remains a consistent bottleneck despite strong performance on local semantic tasks; (ii) similar overall accuracy can correspond to substantially different capability profiles; and (iii) overconfident errors on difficult items concentrate on logically constrained reasoning, highlighting reliability risks that are not captured by standard evaluation metrics.[1]

1 Introduction

Large language models (LLMs) have achieved remarkable progress across a wide range of natural language understanding tasks, including question answering, reading comprehension, and educational applications (Xiao et al., 2023; Cheng et al., 2024; Dong et al., 2024). Along with these advances, increasing attention has been devoted to applying LLMs to classical text processing and digital humanities research (Winata et al., 2024; Yao et al., 2025b). Among such tasks, Literary Chinese reading comprehension remains particularly challenging due to the languages condensed expressions, flexible syntax, frequent omissions, and strong reliance on contextual inference (Zhang et al., 2024). Unlike modern Chinese, Literary Chinese often lacks explicit grammatical and semantic cues, requiring models to recover implicit meanings and discourse relations from limited surface information, thereby placing higher demands on semantic understanding and reasoning capabilities (Cahyawijaya et al., 2024).

To evaluate LLM performance on this task, a series of benchmarks have recently been introduced, including CCLUE (Xu et al., 2020), ACRE (Rao et al., 2023), WYWEB (Zhou et al., 2023), AC-EVAL (Wei et al., 2024), E-EVAL (Hou et al., 2024), and CRISIS (Rao et al., 2025). While these benchmarks provide valuable testbeds for Literary Chinese understanding, their evaluations are still largely based on aggregate accuracy, focusing primarily on whether a model answers correctly. However, Literary Chinese reading comprehension involves diverse cognitive demands, such as modern–classical sense disambiguation, fact localization, conditional reasoning, and cross-paragraph discourse integration. Treating all items as homogeneous therefore makes it difficult to reveal which capabilities models truly master and where their weaknesses originate.

To address this limitation, recent studies have introduced Item Response Theory (IRT) into LLM evaluation, aiming to model latent abilities together with item characteristics such as difficulty and discrimination (Liu et al., 2023; Zhao et al.; Zhou et al., 2026). Compared with aggregate metric, IRT provides a more fine-grained perspective on model performance by distinguishing whether failures arise from limited model ability or intrinsically challenging items. However, ability estimation alone is still insufficient for understanding

[1]Code will be available at https://github.com/JJoinW/DELit.
[†]These authors contributed equally to this work.
[*]Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 12991–13005
October 24-29, 2026 ©2026 Association for Computational Linguistics

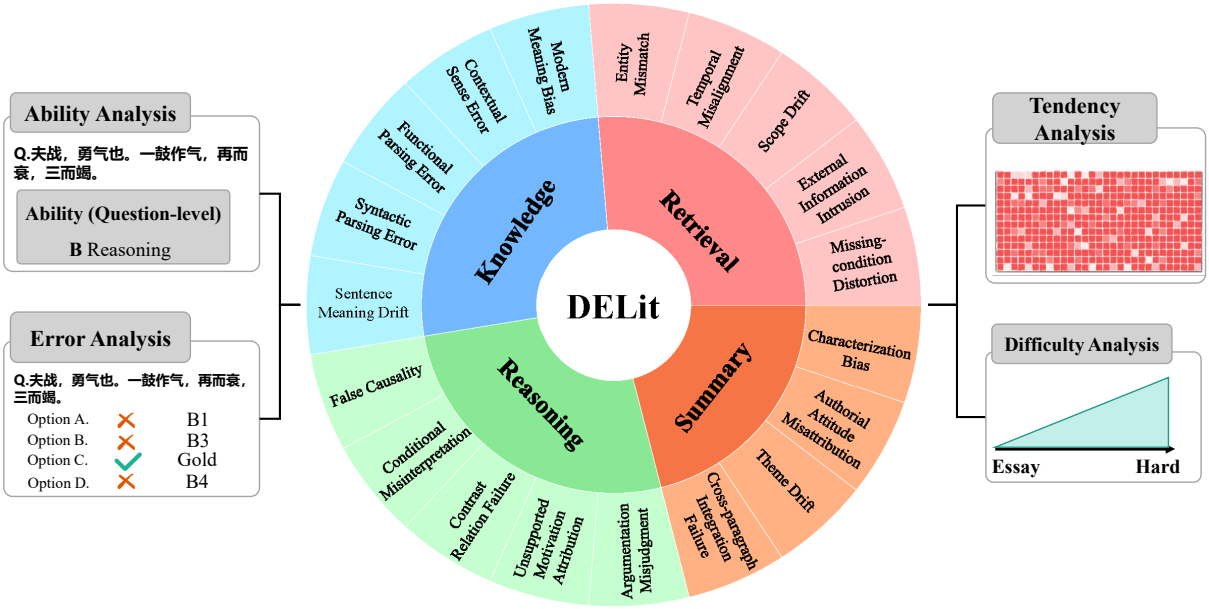


Figure 1: The DELit capability–error diagnostic framework for Literary Chinese reading comprehension.

model behavior in Literary Chinese reading comprehension. Although IRT can infer that a model lacks ability on certain items, it cannot explain why a particular distractor is selected or what kind of reasoning failure underlies the prediction. In practice, errors may stem from modern–classical sense confusion, incorrect causal inference, or failures in cross-paragraph discourse integration, all of which require more detailed diagnostic analysis beyond latent ability measurement alone.

In this paper, we propose **DELit** (Diagnostic Evaluation for Literary Chinese reading comprehension), a diagnostic evaluation framework for Literary Chinese reading comprehension. Unlike traditional evaluations that mainly rely on aggregate metrics such as accuracy, DELit aims to further reveal what capabilities LLMs possess and what types of errors they are prone to in low-resource Literary Chinese reading scenarios. As illustrated in Figure 1, DELit treats LLM responses as observable manifestations of underlying reading abilities, and integrates ability estimation, error attribution, and reliability analysis into a unified evaluation framework. Specifically, we first construct a hierarchical *capability–error taxonomy* for Literary Chinese reading comprehension, where item-level abilities are organized into four dimensions: Literary Chinese knowledge understanding, factual alignment, implicit relation reasoning, and discourse-level modeling, together with 19 fine-grained error types annotated at the option level. We then introduce a *Deep-MIRT* model to jointly estimate latent model abilities, item difficulty, and item discrimination, enabling fine-grained analysis of model performance across different reading capabilities. Based on this framework, we further perform capability-decoupled residual analysis and high-confidence error analysis to uncover stable vulnerabilities and hidden reliability risks that cannot be captured by aggregate evaluation scores alone.

The main contributions of this paper are as follows:

1. We construct a hierarchical capability–error annotation scheme for Literary Chinese reading comprehension, enabling fine-grained analysis at both ability and error levels.

2. We propose DELit, a diagnostic evaluation framework that integrates capability-aware Deep-MIRT modeling with residual-based reliability analysis for fine-grained LLM evaluation.

3. We conduct extensive experiments on mainstream LLMs and reveal several hidden weaknesses and reliability risks that cannot be captured by aggregate metric alone.

2 Related Work

Literary Chinese benchmarks. Literary Chinese, also known as Classical or Ancient Chinese, constitutes an important part of Chinese cultural heritage (Daddo, 2024). However, due to

its low-resource nature and distinctive linguistic characteristics, it remains challenging for modern large language models (LLMs) (Zhang et al., 2024; Cahyawijaya et al., 2024). To evaluate LLM performance on Literary Chinese understanding, a series of benchmarks and datasets have recently been proposed. Existing resources include ACRE (Rao et al., 2023) and Native Chinese Reader (Xu et al.), which focus on Literary Chinese reading comprehension, as well as broader evaluation benchmarks such as WYWEB (Zhou et al., 2023), AGIEval (Zhong et al., 2024), E-EVAL (Hou et al., 2024), AC-EVAL (Wei et al., 2024), and the integrated benchmark proposed by (Cao et al., 2024). These benchmarks have substantially expanded the evaluation coverage of Literary Chinese understanding tasks, ranging from factual comprehension to reasoning and discourse analysis. Despite these advances, existing evaluations are still largely centered on aggregate metrics such as accuracy. While such metrics provide an overall measure of model performance, they offer limited insight into the underlying cognitive abilities required by different items or the specific error patterns exhibited by models.

IRT-based LLM evaluation. Recent studies have shown that aggregate evaluation metrics are highly sensitive to benchmark composition and item difficulty, and therefore may not reliably reflect the actual capabilities of LLMs (Sun et al., 2024; Liu et al., 2025; Xiang et al., 2026). To address this issue, Item Response Theory (IRT) (Lord and Novick, 2008), a psychometric framework for jointly modeling test takers and test items, has increasingly been introduced into LLM evaluation (Castleman et al., 2025; Yao et al., 2025a; Zhou et al., 2026). Existing work has applied IRT to improve the validity of benchmark comparisons, analyze latent model abilities, and reduce the influence of dataset difficulty (Castleman et al., 2025; Liu et al., 2023; Zhou et al., 2026). Despite these advances, most existing IRT-based evaluations mainly focus on latent ability estimation while providing limited insight into the fine-grained cognitive skills and error behaviors underlying model predictions. In addition, item abilities are often modeled in a relatively homogeneous manner without explicitly considering the diverse capability requirements of complex reading comprehension tasks. In contrast, DELit combines Deep-MIRT with a hierarchical capability–error taxonomy, enabling capability-aware ability estimation and fine-grained diagnostic analysis of model errors.

3 The DELit Framework

3.1 Overview

As illustrated in Figure 2, DELit consists of three stages for diagnostic evaluation of Literary Chinese reading comprehension. Stage 1 (ğ3.2) constructs a hierarchical capability–error taxonomy for Literary Chinese reading comprehension. Each question is annotated with its required reading capabilities, while each distractor option is further associated with a specific error type. This process transforms conventional accuracy-based evaluation into fine-grained capability- and error-level analysis. Stage 2 (ğ3.3) introduces a capability-aware Deep-MIRT model to jointly estimate latent model abilities together with item difficulty and discrimination. By incorporating Q-matrix constraints derived from the capability annotations, this stage enables fine-grained measurement of model performance across different reading abilities. Stage 3 (ğ3.4) further performs residual-based diagnostic analysis by comparing observed model behaviors with the predictions of Deep-MIRT. This stage identifies stable error tendencies and reliability risks that cannot be explained solely by latent ability or item difficulty.

3.2 Capability–Error Taxonomy

We build upon CRISIS (Rao et al., 2025), a Literary Chinese benchmark specifically designed for passage-level reading comprehension. Unlike datasets that rely heavily on historical fact memorization, CRISIS focuses on questions requiring contextual understanding and reasoning over passages, making it more suitable for fine-grained diagnostic evaluation of reading abilities. The benchmark provides "passage–question–option" tuples for Literary Chinese reading comprehension, but its original annotations mainly support binary correctness evaluation.

To enable fine-grained analysis, we further re-annotate all 4,417 items with a hierarchical capability–error taxonomy consisting of 4 capability dimensions and 19 fine-grained error types.

Capability labels are assigned at the item level to characterize the primary reading ability required for solving a question, while error labels are assigned at the option level to describe the

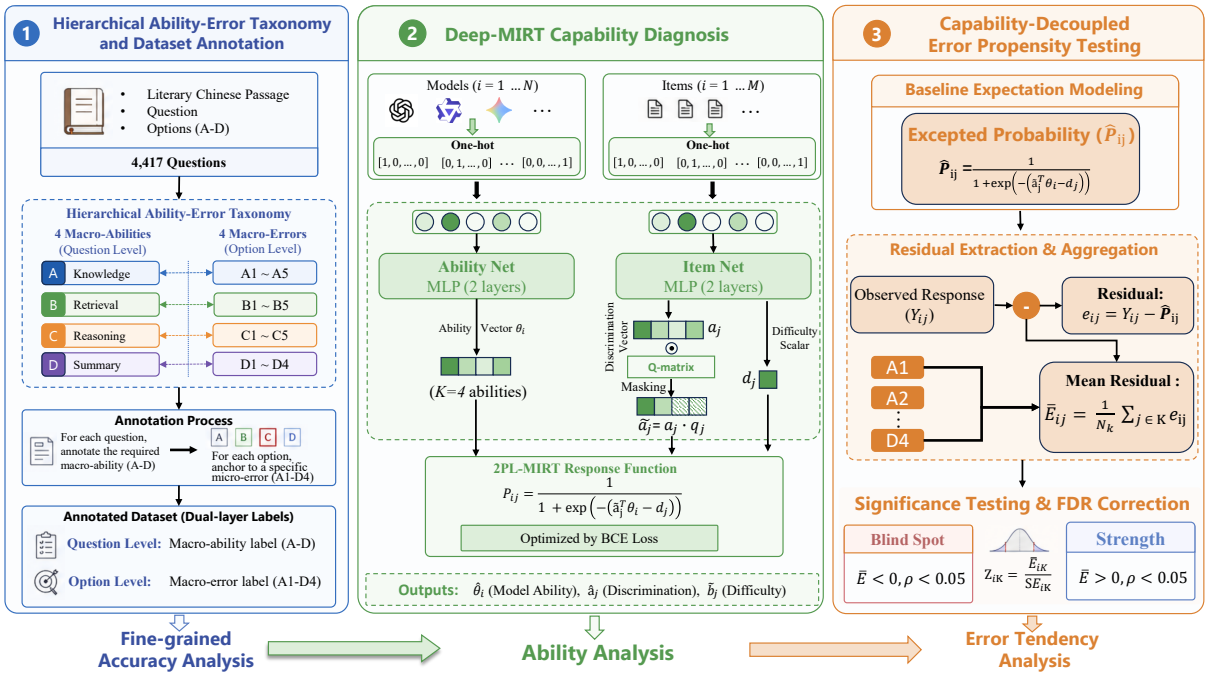


Figure 2: Overall framework of DELit. (1) A hierarchical ability–error taxonomy is used to annotate CRISIS at both the question (ability) and option (error) levels. (2) A Deep-MIRT model maps LLMs and items into continuous latent parameter spaces while preserving MIRT interpretability through the Q-matrix and a 2PL response function. (3) Residual-based significance testing with FDR correction identifies cognitive propensities disentangled from ability, item difficulty, and data distribution.

specific misunderstanding or reasoning pattern reflected by each distractor. In addition, each error type is associated with a corresponding capability dimension, enabling more interpretable analysis of model behaviors beyond correct-or-incorrect evaluation. Table 1 summarizes the overall taxonomy.

GPT-5.4 was used to annotate the full dataset according to the predefined taxonomy, and two human annotators independently re-annotated a randomly sampled 10% subset without access to the GPT-generated labels. Human–human Cohen's κ was 0.79 for capability labels and 0.71 for error labels; after adjudication, GPT–human agreement was 0.82 and 0.73, respectively. More annotation details and label distributions are provided in Appendix A.

3.3 Deep-MIRT Ability Estimation

Although capability-level accuracy provides a more fine-grained evaluation view, it still cannot account for differences in item difficulty and discrimination. We therefore adopt a Multidimensional Item Response Theory (MIRT) to estimate latent reading abilities of LLMs.

For N models and M items, we encode model and item IDs as one-hot vectors and map them through two-layer MLPs into continuous latent parameters (Zhou et al., 2026): an ability vector $\boldsymbol{\theta}_i \in \mathbb{R}^K$, a discrimination vector $\mathbf{a}_j \in \mathbb{R}^K$, and a difficulty scalar b_j, with $K = 4$ corresponds to the four capability dimensions defined in our taxonomy.

To incorporate capability annotations into ability estimation, we further introduce a Q-matrix $\mathbf{Q} \in \{0,1\}^{M \times K}$, where each row $\mathbf{q}_j$ indicates the capability requirements of item j. The discrimination vector is masked by the corresponding capability annotation:

$$\tilde{\mathbf{a}}_j = \mathbf{a}_j \odot \mathbf{q}_j, \tag{1}$$

where $\odot$ denotes element-wise multiplication. The probability that model i correctly answers item j is then computed as:

$$P_{ij} = \sigma\left(\tilde{\mathbf{a}}_j^\top \boldsymbol{\theta}_i - b_j\right). \tag{2}$$

We use a compensatory 2PL form (see Appendix H for details) and train end-to-end with binary cross-entropy on the observed correctness $y_{ij} \in \{0,1\}$. The neural component only searches latent parameters, preserving interpretability and yielding per-model abilities, per-item difficulties,

Capability	Error Type	Definition
Classical Chinese Knowledge	Modern Meaning Bias	Interpreting a Literary Chinese word with its modern sense, ignoring historical semantics and context.
	Contextual Sense Error	Failing to select the correct sense given the context, yielding local semantic errors.
	Functional Parsing Error	Misjudging part-of-speech conversion, function-word usage, or grammatical role.
	Syntactic Parsing Error	Misanalyzing Literary Chinese constructions such as inversion and ellipsis.
	Sentence Meaning Drift	Word-level reading is plausible, but the sentence interpretation drifts from the original.
Text–Fact Alignment	Entity Mismatch	Confusing persons, subjects, objects, or recipients of actions.
	Temporal Misalignment	Misordering events or misreading their temporal stage.
	Scope Drift	Expanding, narrowing, or exaggerating quantity, scope, or degree.
	External Information Intrusion	Introducing information, assumptions, or evaluations unsupported by the text.
	Missing-condition Distortion	Ignoring a key condition or restriction, thereby altering the factual claim.
Implicit Relation Inference	False Causality	Inferring temporal adjacency or coordination as causal.
	Conditional Misinterpretation	Treating hypotheses or unrealized conditions as accomplished facts.
	Contrast Relation Failure	Missing contrast, concession, or progression.
	Unsupported Motivation Attribution	Assigning intentions or motives not supported by the text.
	Argumentation Misjudgment	Misreading analogy, contrast, exemplification, or argument structure.
Discourse-Level Modeling	Characterization Bias	Over-generalizing a character's traits from local evidence.
	Authorial Attitude Misattribution	Misjudging the author's or narrator's stance or evaluation.
	Theme Drift	Mistaking local content for the central theme, or overinterpreting it.
	Cross-paragraph Integration Failure	Failing to coherently integrate information across paragraphs.

Table 1: Capability dimensions and error types for Literary Chinese reading comprehension.

and a calibrated baseline expectation used by the residual test below.

3.4 Capability-Decoupled Error Tendency Test

Although Deep-MIRT models latent abilities together with item difficulty and discrimination, models may still exhibit systematic preferences or weaknesses toward particular error types. To further analyze such behaviors, we perform residual-based error tendency analysis on top of Deep-MIRT predictions.

Let $\hat{P}_{ij}$ denote the probability predicted by Deep-MIRT for model i on item j, and let $Y_{ij} \in \{0, 1\}$ denote the observed response correctness. For each error category K, let $\mathcal{I}_K$ denote the set of items containing at least one distractor annotated with K, and let $N_K = |\mathcal{I}_K|$; these category-specific sets may overlap, but each item is counted only once within a category. We compute the average residual as:

$$\bar{E}_{iK} = \frac{1}{N_K} \sum_{j \in \mathcal{I}_K} \left(Y_{ij} - \hat{P}_{ij} \right). \tag{3}$$

A positive residual indicates that the model performs better than expected on a specific error category given its estimated ability, while a negative residual indicates worse-than-expected performance. We further conduct statistical significance testing on residual scores and apply Benjamini–Hochberg FDR correction at $\alpha = 0.05$ across all model–error pairs. Significant residual patterns therefore reveal stable behavioral tendencies that cannot be fully explained by latent ability alone.

4 Experiments

To systematically investigate the limitations of aggregate evaluation in Literary Chinese reading comprehension, we design three research questions that progressively move from capability structure, to residual error behavior, and finally

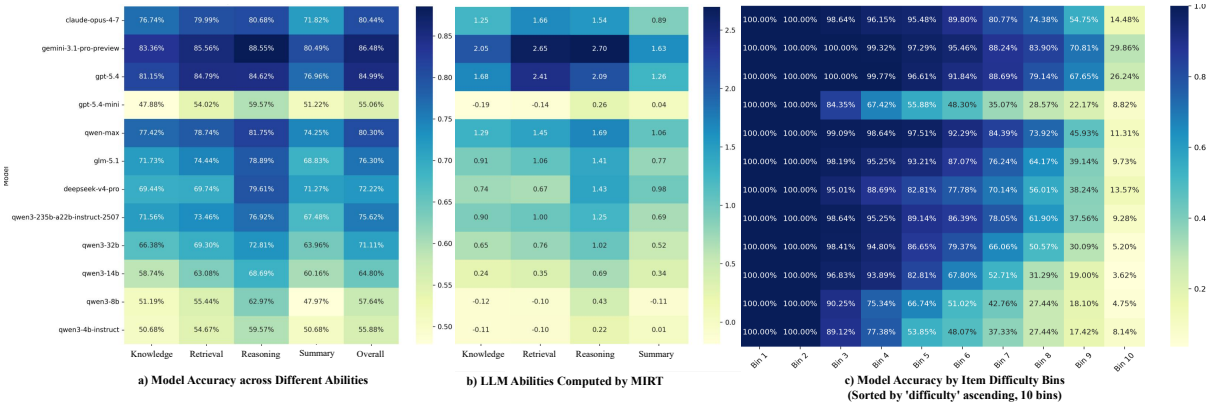


Figure 3: Capability diagnosis of LLMs on Literary Chinese reading comprehension. (a) Accuracy across the four capability dimensions. (b) Deep-MIRT latent ability estimates. (c) Accuracy across Deep-MIRT difficulty bins.

to reliability under uncertainty. Specifically, RQ1 examines whether aggregate accuracy obscures underlying capability differences across models; RQ2 investigates whether models exhibit stable, capability-independent error preferences after controlling for latent ability; and RQ3 focuses on high-confidence failure cases to analyze the most critical reliability risks in deployment settings.

4.1 Setup

Data. We conduct experiments on the 4,417 annotated CRISIS items described in ğ3.2, covering four capability dimensions and 19 error types.

Models. We evaluate a diverse set of mainstream LLMs, including GPT, Gemini, Claude, DeepSeek, and Qwen series models, under a unified zero-shot multiple-choice setting. Model outputs are converted into a binary response matrix $\mathbf{Y} \in \{0,1\}^{N\times M}$ for Deep-MIRT estimation.

Implementation. Deep-MIRT jointly estimates latent model abilities and item characteristics based on $\mathbf{Y}$ using the formulation in Eq. (2). Residual-based error tendency analysis is then performed on top of the estimated response probabilities. The selected compensatory 2PL model achieves an ACC of 0.8246, an F1 score of 0.8812, and an AUC of 0.8734 on held-out response prediction; implementation details and model-selection results are reported in Appendices B and H, respectively.

4.2 RQ1: Does aggregate metric mask capability structure?

Figure 3(a,b) compares ability accuracy with Deep-MIRT latent ability estimates, while Figure 3(c) presents performance under different item difficulty levels. The results reveal several patterns that are not observable from aggregate accuracy alone.

Models exhibit heterogeneous capability profiles. Although some models achieve similar overall accuracy, their performance distributions across capability dimensions differ substantially. In particular, several models achieve relatively strong performance on *Knowledge* and *Retrieval* items but show noticeably weaker results on *Summary*. By contrast, capability differences on *Reasoning* are comparatively smaller across models. These findings suggest that aggregate accuracy may conceal important differences in underlying reading abilities.

Capability distributions vary across model families. Within the Qwen3 family, capability improvements remain relatively consistent across all four dimensions as parameter scale increases (Appendix C). However, different model families exhibit distinct capability tendencies. For example, Gemini-series models generally perform better on *Retrieval*, while GPT-series models show relatively stronger performance on *Knowledge* and *Reasoning*. These results indicate that similar overall accuracy can correspond to substantially different capability structures.

Performance differences become larger on difficult items. Figure 3(c) shows that model performance decreases consistently as item difficulty increases. On easy items, most models achieve similarly high accuracy, whereas performance differences become much larger on difficult items. Moreover, the ranking consistency between aggregate accuracy and Deep-MIRT ability estimates decreases on high-difficulty subsets, suggesting

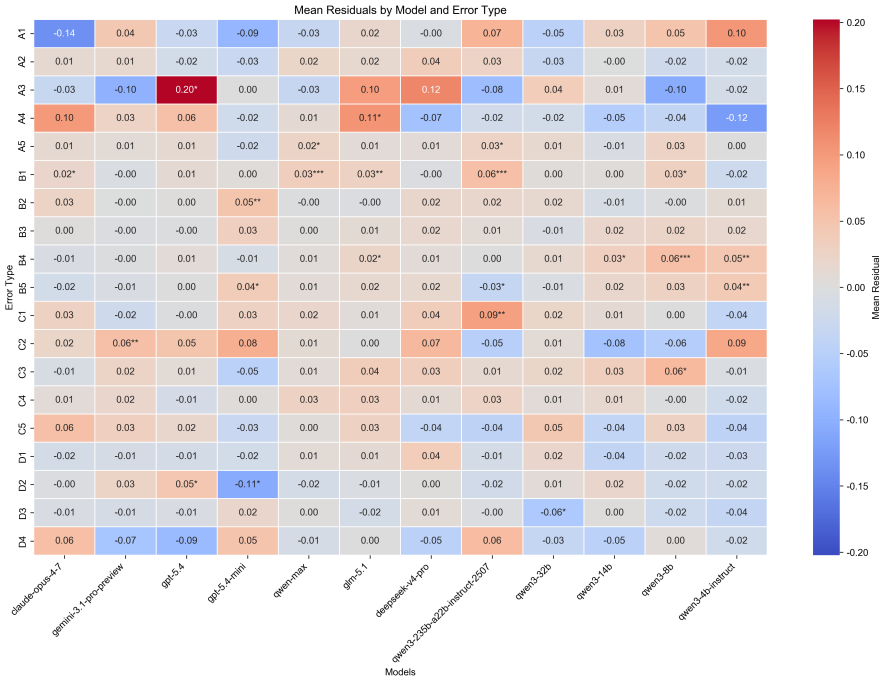


Figure 4: Residual matrix between Deep-MIRT predicted probabilities and observed responses across 19 error types. Values are aggregated item-level residuals and corrected using FDR ($\alpha = 0.05$). Positive (negative) values indicate better (worse) performance than expected under the estimated latent ability model.

that difficult items are more informative for distinguishing latent reading abilities. Additional analyses on difficult items are provided in Appendix D.

Overall, the results demonstrate that aggregate metric alone is insufficient for understanding LLM performance on Literary Chinese reading comprehension. Fine-grained capability analysis and difficulty-aware evaluation provide a more informative view of model strengths and weaknesses.

4.3 RQ2: After controlling for ability, do models show stable error preferences?

Figure 4 presents the residual-based error tendency analysis after controlling for latent ability, item difficulty, and discrimination through Deep-MIRT. Although most residual values remain close to zero, several model–error pairs still exhibit statistically significant deviations after FDR correction, indicating stable error preferences beyond overall ability estimation.

Certain models show targeted robustness on specific error types. Positive residuals are relatively sparse but reveal interpretable behavioral patterns. For example, *gpt-5.4* performs significantly better than expected on *A3 (Functional Parsing Error)*, suggesting relatively strong robustness in handling function-word usage and part-of-speech conversion. Similarly, *gemini-3.1-pro-preview* shows positive residuals on *C2 (Conditional Misinterpretation)*, indicating stronger-than-expected performance on conditional reasoning. Notably, these positive residuals are concentrated on specific error categories rather than consistently appearing across related categories, suggesting that model robustness is often localized rather than uniformly distributed.

Negative residuals concentrate on discourse-level constraint maintenance. Compared with positive residuals, negative residuals are more common and mainly appear in *B5 (Missing-condition Distortion)*, *D2 (Authorial Attitude Misattribution)*, and *D3 (Theme Drift)*. These error categories require models to maintain discourse-level constraints, attitudes, or themes across longer contexts rather than relying solely on local fact matching. Importantly, several models with relatively strong overall capability estimates still exhibit significant negative residuals on these categories, suggesting that strong local reasoning ability does not necessarily transfer to globally constrained discourse understanding.

Models behave differently on explicit and implicit fact-alignment errors. A notable contrast appears between *B4 (External Information Intru-*

sion) and *B5 (Missing-condition Distortion)*. Most models exhibit near-zero residuals on B4, while significantly negative residuals remain common on B5. Although both categories are related to factual alignment, they involve different reasoning behaviors. B4 mainly reflects explicit fabrication of unsupported information, whereas B5 requires maintaining conditional or scope constraints during multi-step reasoning. These findings suggest that current LLMs are relatively robust against explicit unsupported generation but remain vulnerable to implicit constraint loss during discourse-level inference.

Similar overall accuracy can conceal different error tendencies. Finally, we observe that models with very similar aggregate accuracy may still exhibit substantially different residual patterns across error categories. In several cases, models with nearly identical overall scores diverge significantly on categories such as A3, B5, and D2. This suggests that aggregate evaluation scores alone may be insufficient for characterizing model reliability and behavioral differences in Literary Chinese reading comprehension.

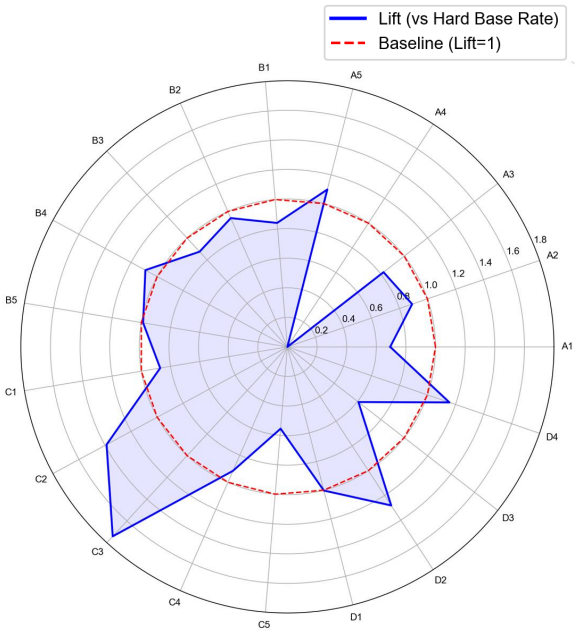


Figure 5: Error-type lift for high-difficulty items with high-confidence incorrect predictions from Qwen3-235B. Lift is defined as the relative frequency compared to the full dataset baseline. Values >1 indicate over-representation.

4.4 RQ3: When models are confident but wrong, what fails?

RQ1 and RQ2 analyze whether models fail and how such failures distribute across capabilities and error types. RQ3 focuses on a more operational setting: cases where models make incorrect predictions with high confidence.

Confidence and difficulty identify but do not eliminate failure cases. Option-level cross-entropy and Deep-MIRT difficulty are both strong predictors of correctness, with AUC values ranging from 0.80 to 0.91 across models (details in Appendix B). As expected, difficult items tend to be associated with higher loss, while easy items show lower loss. However, we observe a small subset of instances that remain incorrectly predicted even under high-confidence and high-difficulty conditions. These cases are not effectively filtered by either difficulty-based or confidence-based criteria, and therefore represent a particularly informative failure subset.

Confident errors concentrate on logical relation reasoning. Figure 5 reports the error-type distribution of this subset relative to the full dataset for Qwen3-235B, selected as a representative open-source model with accessible option-level confidence scores. Two categories are most prominent: *C2 (Conditional Misinterpretation)* and *C3 (Contrast Relation Failure)*. Both require modeling logical relations across clauses rather than relying on local lexical cues.

In these cases, the model often captures surface-level markers of conditional or contrastive structures, leading to low predictive uncertainty, but fails to correctly propagate the corresponding logical constraints during reasoning, resulting in incorrect answers. This suggests a disconnect between local linguistic pattern recognition and global relational inference.

Implications for reliability analysis. These results indicate that benchmark-level calibration metrics may obscure error patterns by aggregating across heterogeneous item types. In particular, confidence-based evaluation may underestimate risks associated with logically structured errors. A more fine-grained analysis over capability and error dimensions can provide a more precise view of model reliability in Literary Chinese reading comprehension.

5 Conclusion

We presented DELit, a diagnostic evaluation framework for Literary Chinese reading comprehension that moves beyond aggregate accuracy toward capability estimation, error attribution, and reliability analysis. Across a range of mainstream LLMs, our analysis shows that aggregate performance measures can obscure important differences in model behavior. In particular, models with similar accuracy may exhibit distinct capability profiles, and their failures concentrate in specific types of discourse-level and logical reasoning challenges. These findings suggest that fine-grained, capability-aware evaluation provides a more informative perspective on LLM behavior in reading comprehension tasks, especially in low-resource and linguistically complex settings such as Literary Chinese.

Limitations

This work has several limitations that suggest directions for future research.

First, our capability–error taxonomy is constructed based on expert linguistic analysis of Literary Chinese reading comprehension. While it covers major error patterns such as lexical mismatch, relation misinterpretation, and discourse-level inconsistency, it may not fully capture more subtle phenomena such as pragmatic inference, cultural allusion, or rhetorical structure understanding. Extending the taxonomy with more fine-grained or data-driven discovery methods is an interesting direction for future work.

Second, the residual-based analysis in DELit is built on top of Deep-MIRT and provides a statistical view of model deviations from expected behavior. However, it does not directly explain internal reasoning processes or representations within LLMs. As a result, the identified error tendencies should be interpreted as behavioral patterns rather than mechanistic explanations. Incorporating interpretability methods or process-level supervision could help bridge this gap.

Finally, our experiments are conducted in a multiple-choice setting, which enables controlled capability estimation and error alignment, but does not fully reflect open-ended generation scenarios. Extending DELit to generative tasks would help evaluate the robustness and generality of the proposed diagnostic framework in more realistic settings.

Ethical Statement

This work uses publicly available Literary Chinese reading comprehension datasets and does not involve any personal or sensitive user data.

The proposed DELit framework is designed for research-oriented diagnostic analysis of large language models, with the goal of understanding capability boundaries and error patterns rather than supporting automated decision-making in high-stakes educational settings. We explicitly caution against using the derived capability estimates or error diagnoses as standalone indicators for evaluating human learners or for automated grading without human oversight.

In particular, while DELit can reveal systematic model weaknesses such as misunderstanding, overconfidence, and sensitivity to implicit constraints, these signals reflect model behavior under controlled evaluation conditions and should not be directly interpreted as pedagogical judgments. Any deployment in educational or assistive scenarios should ensure human-in-the-loop review and transparency regarding known failure modes of the models.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (No. 62477002). We also thank the reviewers for their insightful comments. During the preparation of this work, the authors used generativeAI tools to assist in writing part of the experimental code and for grammar checking.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433.

Jiahuan Cao, Dezhi Peng, Peirong Zhang, Yongxin Shi, Yang Liu, Kai Ding, and Lianwen Jin. 2024. Tonggu: Mastering classical chinese understanding with knowledge-grounded large language models.

In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4196–4210.

Jane Castleman, Nimra Nadeem, Tanvi Namjoshi, and Lydia T Liu. 2025. Rethinking math benchmarks for llms using irt. *Proceedings of Machine Learning Research*, 273:66–82.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2024. Adapting large language models via reading comprehension. In *International Conference on Learning Representations*, volume 2024, pages 48624–48652.

Emily Daddo. 2024. An introduction to ancient china. *Teaching History*, 58(3):67–69.

Bingyu Dong, Jie Bai, Tao Xu, and Yun Zhou. 2024. Large language models in education: A systematic review. In *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, pages 131–134. IEEE.

Jinchang Hou, Chang Ao, Haihong Wu, Xiangtao Kong, Zhigang Zheng, Daijia Tang, Chengming Li, Xiping Hu, Ruifeng Xu, Shiwen Ni, et al. 2024. E-eval: A comprehensive chinese k-12 education evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7753–7774.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yan Liu, Renren Jin, Ling Shi, Zheng Yao, and Deyi Xiong. 2025. Finemath: A fine-grained mathematical evaluation benchmark for chinese large language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(12):1–15.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2023. What we evaluate when we evaluate recommender systems: Understanding recommender systems performance using item response theory. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 658–670.

Frederic M Lord and Melvin R Novick. 2008. *Statistical theories of mental test scores*. IAP.

Dongning Rao, Guanju Huang, and Zhihua Jiang. 2023. Ancient chinese machine reading comprehension exception question dataset with a non-trivial model. In *Pacific Rim International Conference on Artificial Intelligence*, pages 145–158. Springer.

Dongning Rao, Rongchu Zhou, Peng Chen, and Zhihua Jiang. 2025. A comprehensive literary chinese reading comprehension dataset with an evidence curation based solution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3583–3603.

Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. 2024. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19053–19061.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Yuting Wei, Yuanxing Xu, Xinru Wei, Yangsimin Yangsimin, Yangfu Zhu, Yuqing Li, Di Liu, and Bin Wu. 2024. Ac-eval: Evaluating ancient chinese language understanding in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1600–1617.

Genta Indra Winata, Ruochen Zhang, and David Ifeoluwa Adelani. 2024. Miners: Multilingual language models as semantic retrievers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2742–2766.

Kun Xiang, Zhili Liu, Terry Jingchen Zhang, Yinya Huang, Yunshuang Nie, Kaixin Cai, Yiyang Yin, Runhui Huang, Hanhui Li, Yihan Zeng, et al. 2026. Atomthink: Multimodal slow thinking with atomic step reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Changrong Xiao, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Lei Xia. 2023. Evaluating reading comprehension exercises generated by llms: A showcase of chatgpt in education applications. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 610–625.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th international conference on computational linguistics*, pages 4762–4772.

Shusheng Xu, Yichen Liu, Xiaoyu Yi, Siyuan Zhou, Huizi Li, and Yi Wu. Native chinese reader: A dataset towards native-level chinese machine reading comprehension. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Louie Hong Yao, Nicholas Jarvis, Tiffany Zhan, Saptarshi Ghosh, Linfeng Liu, and Tianyu Jiang. 2025a. Je-irt: A geometric lens on llm abilities through joint embedding item response theory. *arXiv preprint arXiv:2509.22888*.

Xinyu Yao, Mengdi Wang, Bo Chen, and Xiaobing Zhao. 2025b. Wenyangpt: a large language model for classical chinese tasks. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 8339–8347.

Jinyi Zhang, Ke Su, Haowei Li, Jiannan Mao, Ye Tian, Feng Wen, Chong Guo, and Tadahiro Matsumoto. 2024. Neural machine translation for low-resource languages from a chinese-centric perspective: A survey. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(6):1–60.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. Can we predict performance of large models across vision-language tasks? In *Forty-second International Conference on Machine Learning*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the association for computational linguistics: NAACL 2024*, pages 2299–2314.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. Wyweb: A nlp evaluation benchmark for classical chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319.

Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, et al. 2026. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35085–35093.

A Annotation Protocol and Label Distribution

We annotate the 4,417 CRISIS items in two passes. First, GPT-5.4 produces a capability label per item and an error-type label per distractor option, conditioned on the taxonomy in Table 1 and short definitions. Second, two human annotators independently re-annotate a 10% random sample without access to the GPT-generated labels. Human–human Cohen's κ is 0.79 for capability labels and 0.71 for error labels; after adjudication, agreement between GPT-5.4 and the final human labels is 0.82 and 0.73, respectively. Figure 6 shows the label distribution. Text–fact alignment dominates the capability axis, and the error-type counts are long-tailed, so we report sample sizes alongside residuals and lifts in the main text to avoid over-interpreting low-support categories.

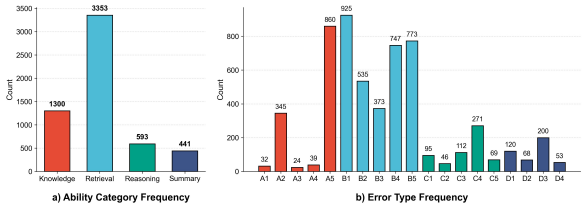


Figure 6: Annotated label distribution. Left: four capability dimensions. Right: 19 error types.

B Implementation Details

B.1 Deep-MIRT Training Details

Deep-MIRT is trained on the binary response matrix $\mathbf{Y} \in \{0,1\}^{N\times M}$ using the objective defined in Eq. (2). We optimize the model with Adam using a learning rate of 10^{-3}, weight decay of 10^{-4}, and batch size of 1024. Training proceeds until convergence on a held-out validation split.

Following prior Deep-IRT studies, we parameterize latent model abilities and item characteristics through neural representations. We experiment with both compensatory and non-compensatory MIRT variants and ultimately adopt the compensatory 2PL formulation in the main experiments due to its superior empirical stability and predictive performance. Additional ablation results are provided in Appendix H.

B.2 Residual Significance Testing

For residual-based error tendency analysis, we first compute the difference between observed responses and Deep-MIRT predicted probabilities for each model–item pair. Residuals are then aggregated according to error categories defined in the capability–error taxonomy.

To identify statistically significant behavioral tendencies, we compute significance scores for each model–error pair and apply Benjamini–Hochberg false discovery rate (FDR) correction with $\alpha = 0.05$ across all comparisons. This procedure controls for multiple testing while preserving sensitivity to stable error patterns.

B.3 Confidence Analysis

To further analyze model reliability, we perform high-confidence error analysis based on option-level prediction distributions. Specifically, we use the cross-entropy loss of the predicted option distribution as a proxy for model uncertainty, where lower cross-entropy corresponds to higher confidence.

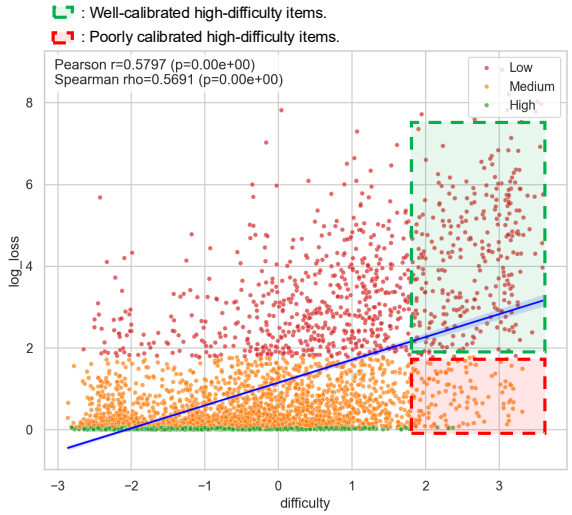


Figure 7: Joint distribution of option-level cross-entropy loss and Deep-MIRT item difficulty for qwen3-235b, illustrating the confidence–difficulty mismatch on the confidently-wrong subset.

We focus on instances that simultaneously exhibit high item difficulty and high-confidence incorrect predictions. These cases reveal situations in which models produce confident but unreliable responses, thereby exposing reliability risks that are not reflected by aggregate evaluation metrics alone.

For each evaluated model, we compute the AUC of (i) option-level negative cross-entropy and (ii) Deep-MIRT item difficulty as predictors of per-item correctness. As an illustrative case, qwen3-235b achieves 0.8033 (confidence) and 0.9099 (difficulty); both signals are predictive but neither alone captures the confidently-wrong subset analyzed in ğ4.4. Figure 7 shows the joint distribution of cross-entropy and difficulty for this model.

B.4 Model Configurations

We evaluate a diverse set of mainstream LLMs spanning both proprietary and open-source families, including GPT-series (Achiam et al., 2023), Gemini-series (Team et al., 2024), Claude-series, DeepSeek-series (Liu et al., 2024), and Qwen-series (Yang et al., 2025). All models are evaluated under a unified zero-shot multiple-choice setting with identical prompts and decoding configurations to ensure fair comparison.

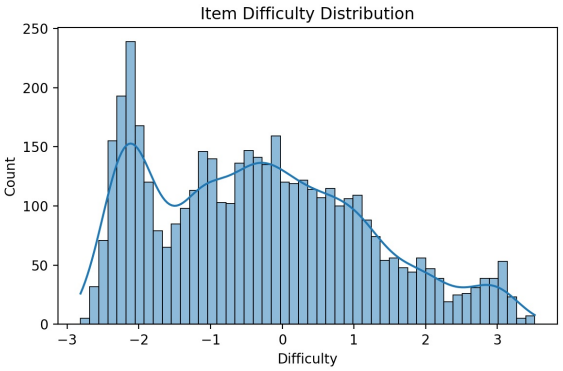


Figure 8: Distribution of item difficulty parameters.

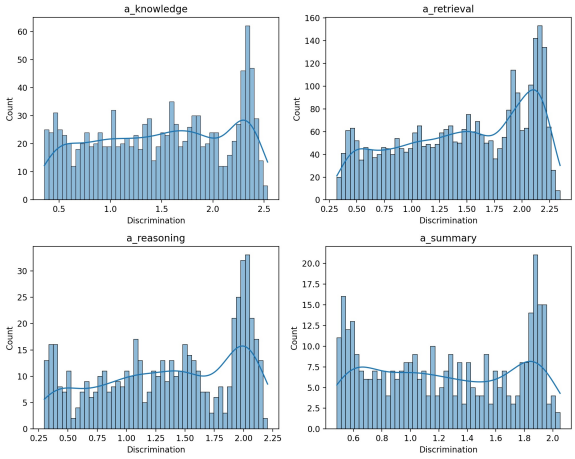


Figure 9: Distribution of item discrimination parameters.

C Deep-MIRT Item Parameter Distributions

Figures 8 and 9 show the global distributions of the difficulty and discrimination parameters estimated by Deep-MIRT. Neither distribution is concentrated on extreme values, supporting the use of difficulty bins in the main text. Figures 10 and 11 further break these parameters down by error type. Error types with simultaneously high difficulty and high discrimination (e.g. several relation and discourse categories) are both harder *and* more capability-discriminative; categories with low discrimination should be interpreted with more caution since residual variation on them is less informative about ability.

D Fine-Grained Error-Type Accuracy on High-Difficulty Items

Figure 12 provides the full per-model accuracy on different error type items, complementing Figure 3. The pattern matches the main text: relation-

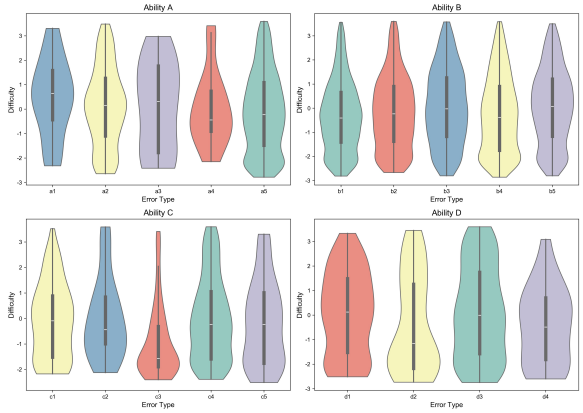


Figure 10: Item difficulty by error type.

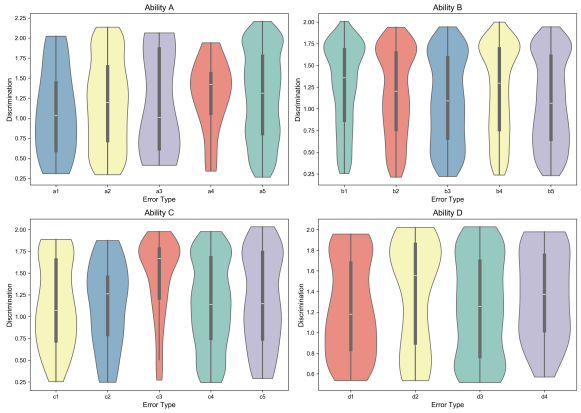


Figure 11: Item discrimination by error type.

type errors C1–C3 are largely handled, A1 (Modern Meaning Bias) and A3 (Functional Parsing Error) drive most of the variance across models, and discourse-level errors D1–D3 remain low across the board.

E B4 vs. B5: Explicit vs. Implicit Constraint Violations

B4 (External Information Intrusion) and B5 (Missing-condition Distortion) are both fact-alignment errors but probe different failure modes. B4 measures whether a model fabricates content not supported by the passage — the canonical "hallucination" setting — whereas B5 measures whether a model can carry conditional and scope constraints through multi-step inference. The residual pattern in ğ4.3 shows that current LLMs have substantially closed the B4 gap but have not made comparable progress on B5, suggesting that the remaining "hallucination-like" behavior in classical reading is increasingly about implicit constraint maintenance rather than explicit external fabrication. We view this as a useful operational distinction for designing supervision signals.

F Lift Formulation

For model m and error type k, let $\mathcal{I}_m$ be the set of items m answered incorrectly and $\mathcal{D}$ the full item set; t_j denotes the error type label of item j. Define

$$p_{m,k}^{\text{wrong}} = \frac{\sum_{j \in \mathcal{I}_m} \mathbf{1}[t_j = k]}{|\mathcal{I}_m|}, \tag{4}$$

$$p_k^{\text{all}} = \frac{\sum_{j \in \mathcal{D}} \mathbf{1}[t_j = k]}{|\mathcal{D}|}, \tag{5}$$

$$\text{Lift}_{m,k} = \frac{p_{m,k}^{\text{wrong}}}{p_k^{\text{all}}}. \tag{6}$$

$\text{Lift}_{m,k} > 1$ means error type k is over-represented among m's mistakes relative to the dataset baseline; < 1 means under-represented. Lift captures shifts in error *composition* and is therefore complementary to absolute error rates. A model-wise lift radar over all error types is shown in Figure 14; Qwen3-family capability radars are shown in Figure 13.

G Residual Standard Error and Significance Test

Following ğ3.4, under the null hypothesis of no specific tendency the mean residual $\bar{E}_{iK}$ has Bernoulli-derived standard error

$$SE_{iK} = \sqrt{\frac{1}{N_K^2} \sum_{j \in \mathcal{I}_K} \hat{P}_{ij}\,(1 - \hat{P}_{ij})}, \tag{7}$$

and we form $Z_{iK} = \bar{E}_{iK} / SE_{iK}$. Because we test all model $\times$ error-type cells jointly, we apply Benjamini–Hochberg FDR correction at $\alpha = 0.05$ and report only the surviving residuals as significant tendencies.

H Deep-MIRT Model Selection and Ablations

Table 2 compares six Deep-IRT variants crossed over compensatory vs. noncompensatory mechanisms and 1PL/2PL/3PL forms, and also reports Kendall τ between estimated abilities and the four capability axes. The compensatory 2PL variant is best on overall metrics and most balanced across capability dimensions. The compensatory advantage is consistent with the nature of Literary Chi-

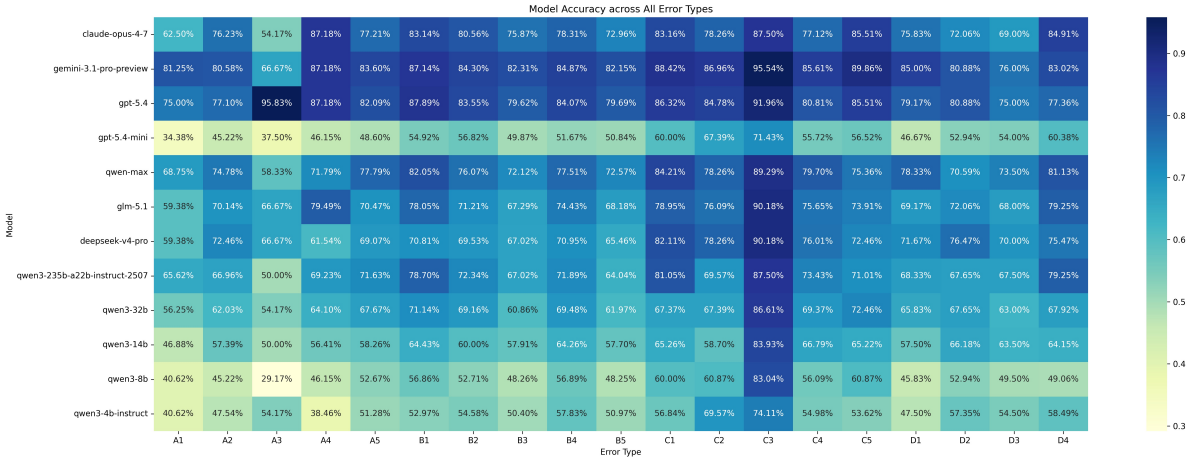


Figure 12: Fine-grained accuracy of LLMs by error type. Darker colors indicate stronger robustness on the corresponding error type.

Model	Mechanism	Params	ACC	F1	AUC	Kendall τ			
						knowledge	retrieval	reasoning	summary
Deep-IRT	Noncompensatory	1PL	0.8180	0.8783	0.8653	0.9091	0.9091	0.8182	0.8788
		2PL	0.8200	0.8755	0.8732	0.9697	0.8788	0.8485	0.9091
		3PL	0.8232	0.8786	0.8724	0.9091	0.9697	0.8788	0.8182
	Compensatory	1PL	0.8161	0.8765	0.8644	0.9394	0.9091	0.8788	0.8182
		2PL	**0.8246**	**0.8812**	0.8721	0.9091	0.9091	**0.9091**	**0.9091**
		3PL	0.8202	0.8757	0.8704	0.9091	0.9091	0.7879	0.8485

Table 2: Deep-IRT variants under compensatory vs. noncompensatory response mechanisms.

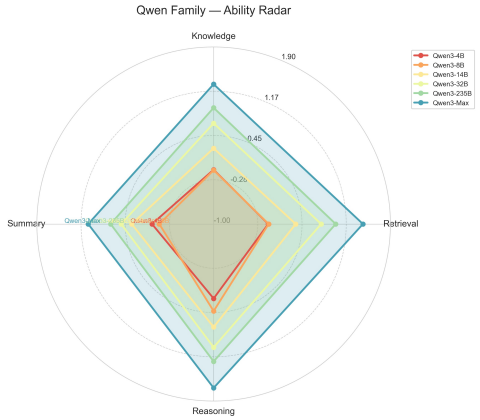


Figure 13: Capability radar of Qwen3-family models on the four capability dimensions.

nese reading, where capabilities such as knowledge and retrieval can partially compensate for weaker discourse-level reasoning. Kendall τ is consistently high for knowledge and retrieval but more variable for reasoning and summary, mirroring the main-text finding that high-order discourse capabilities are the least stable. We therefore use compensatory 2PL as the default Deep-MIRT in the main text.

Table 3 ablates two key designs in Deep-MIRT.

Method	ACC	F1	AUC
Deep-MIRT (full)	**0.8246**	**0.8812**	**0.8734**
w/o Neural Net	0.7804	0.8645	0.8281
w/o Q-Matrix	0.8232	0.8785	0.8721

Table 3: Ablation of Deep-MIRT components.

Removing the neural parameter searcher (*w/o Neural Net*) reduces the model to standard MIRT, dropping ACC from 0.8246 to 0.7804, confirming that the neural component effectively captures nonlinear interactions among models, items, and capability dimensions. Removing the Q-matrix turns the model into a deep-IRT model, with a smaller numerical drop in performance, but loses the structural alignment between items and capabilities, which the rest of DELit (residuals, lift, error-type breakdowns) relies on for interpretability.

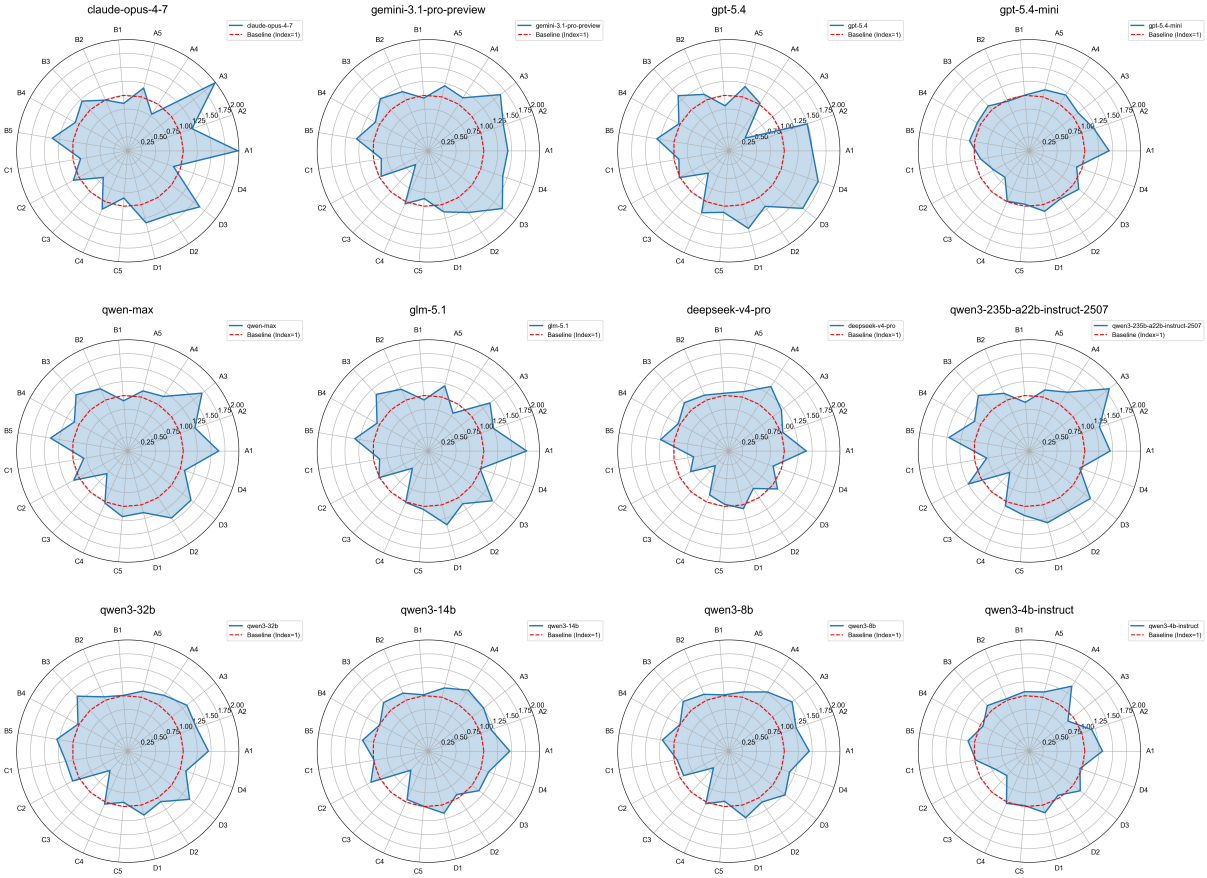


Figure 14: Per-model lift of error types in wrong items relative to the all-item baseline. The red dashed line is Lift $= 1$.