

LLM as Investigator and Judge in Pairwise Comparisons for Item Parameter Modeling

Brian Harrold¹, Vincent Fakiyesi², Maria D'Brot³,
Susan Lottridge³, Justin Barber³, Michael Hemenway³

¹University of South Carolina, ²University of Georgia, ³Pearson

Correspondence: bharrold@email.sc.edu

Abstract

This study synthesizes feature-based and pairwise-comparison approaches to item parameter modeling by using an LLM as an investigator of item characteristics and a judge of relative item parameters. The proposed framework aims to improve recovery of item parameter rankings and predictions while providing insight into factors associated with the parameters.

1 Background

Item parameter modeling refers to the process of predicting Item Response Theory (IRT) parameters, such as difficulty or discrimination, using item characteristics or metadata prior to field testing or operational use. Predicted item parameters can support item screening, field-test form construction, and more efficient calibration procedures (Ulitzsch et al., 2026). Field testing places substantial demands on assessment programs because items cannot be used operationally until they have been administered to sufficient numbers of students, analyzed, and reviewed. More efficient approaches to item parameter modeling may help assessment programs update test designs more quickly, reduce student time spent on non-operational items, and expand item banks to support future assessment needs.

Existing research on item parameter modeling has utilized a wide variety of item characteristics, including linguistic, cognitive, semantic, and content-related features. More recently, advances in large language models (LLMs) have enabled automated generation of item characteristics that extend beyond traditional metadata and hand-crafted features (Peters et al., 2025). Studies have used these features within statistical, machine learning, and transformer-based frameworks to directly predict item parameters. A second line of research approaches item parameter modeling through pair-

wise comparisons. Rather than directly predicting values, judges compare pairs of items and difficulty estimates are rendered through scaling frameworks such as the Bradley-Terry-Luce model (ACER, 2022; Pollitt, 2012). Studies have extended this approach by using LLMs as pairwise judges, demonstrating the potential to recover meaningful item orderings from LLM-based decisions (Li and Gushta, 2026; Han and Rijmen, 2026).

While both approaches have shown promise, they have largely been investigated independently. Direct-prediction approaches focus on identifying and modeling characteristics associated with item difficulty, while pairwise-comparison approaches focus on recovering relative difficulty orderings through comparisons. Little work has examined whether information generated through the first approach can improve performance in the second. This study investigates whether LLM-generated item characteristics and available item metadata can be incorporated into pairwise-comparison methods to improve the recovery of item difficulty rankings while also providing insight into the characteristics that drive perceived item difficulty.

2 Aims

This study examines two complementary roles for LLMs in item parameter modeling. First, the LLM serves as an investigator by characterizing item features, contextual information, and other relevant information. Second, the LLM serves as a pairwise judge by using these characteristics, along with available metadata, to evaluate the relative difficulty of item pairs. The primary focus is to determine whether synthesizing these two roles improves the recovery of empirically derived item rankings compared to other LLM-based approaches, while also identifying LLM-generated features associated with item difficulty.

3 Sample

To evaluate the proposed framework in an operationally relevant setting, this study used Grade 6 English Language Arts (ELA) items from a large-scale statewide assessment program. The items were administered either operationally or as field-test items between Spring 2022 and Spring 2025. Some items were initially field tested and subsequently administered operationally; when multiple administrations were available, psychometric data from the most recent administration were retained for analysis. Items associated with video-based stimuli were excluded because the stimuli could not be represented equivalently within the LLM-based modeling framework, resulting in the exclusion of 29 items. The final analytic sample consisted of 333 items.

The 333 items were associated with 78 unique passages, including 56 substantive reading passages and 22 setup or scenario passages that provided additional contextual information. Most items (91.3%) were associated with a single substantive passage, while 27 items (8.1%) were associated with two substantive passages and two items (0.6%) were associated with three. The item bank included 252 Evidence-Based Selected Response (EBSR) items (75.7%) and 81 Technology-Enhanced Constructed Response (TECR) items (24.3%). EBSR items used standard multiple choice interactions, whereas TECR items represented a range of technology-enhanced response formats, including gap-match, inline-choice, matching, and hybrid interactions.

All items were worth two score points and had been calibrated using the generalized partial credit model (GPCM; Muraki, 1992). Because the items had previously been administered operationally or through field testing, the empirical item parameters were estimated from substantial samples of student responses. The number of students reaching each item ranged from 1,782 to 126,867 (median = 1,976). Empirical item difficulty estimates ranged from -1.809 to 5.275 ($M = 0.422$, $SD = 0.939$), with a median of 0.299 ; the distributions of the empirical item parameters are shown in Figure 1. These empirical difficulty estimates served as the criterion against which the LLM-based pairwise judgments and resulting item difficulty rankings were evaluated.

4 Methods

Available assessment metadata were reviewed to retain information that would be available prior to field testing and was relevant to item difficulty. The retained metadata described characteristics of both passages and items, including passage structure and readability, genre and topic, item and interaction type, cognitive complexity, scoring design, and response structure. Fields containing no usable variation or duplicating other information were excluded. In addition to metadata, LLM-generated features were developed from potential drivers of Grade 6 ELA item difficulty identified by experienced ELA assessment professionals. Two subject matter experts (SMEs) participated in semi-structured interviews, and six additional SMEs provided responses to a survey used to refine and expand the resulting framework. Interview findings were analyzed using reflexive thematic analysis following Braun and Clarke (2006) and Braun and Clarke (2021). Characteristics already represented objectively in the assessment metadata, such as passage length, readability, genre, and item format, were excluded to avoid redundancy. The resulting framework included 29 features representing qualitative aspects of passages, items, responses, and their relationships. Depending on the characteristic, features were operationalized as either ordinal ratings (from 1 to 5) or binary indicators using pre-defined definitions and evaluation criteria. Each feature was evaluated for each item using Claude Opus 4.8. The model was only provided with the item, passage, and response options and definitions or evaluation criteria associated with the features, not any information about student responses or empirical item difficulty. This work was motivated by prior research demonstrating that item characteristics and LLM-generated information can provide useful evidence about item difficulty and help identify characteristics that may contribute to difficulty (Tran et al., 2025; Han et al., 2025; Razavi and Powers, 2026; Baldwin et al., 2021).

Pairwise judgments were obtained using Claude Opus and Qwen3-32B, extending comparative-judgment approaches that have been used with human judges, and more recently, LLM-based judges (Pollitt, 2012; ACER, 2022; Han and Rijken, 2026). A pilot study first evaluated four prompting strategies: Item-Only, Item+Metadata, Item+Features, and Item+Metadata+Features. Providing additional information did not improve con-

ditional pairwise accuracy or correspondence between Bradley–Terry scores and empirical difficulty relative to Item-Only in Claude Opus, although providing more information reduced item-order disagreement. The full analysis therefore retained Item-Only as a baseline and Item+Metadata+Features (hereafter, Item+Both) as the test condition. For each item pair, the model was asked to identify which item it expected to be more difficult for students without being provided any student responses or empirical parameters. The full design produced 55,278 ordered pairwise comparisons. Because LLM-based pairwise judgments can be sensitive to presentation order (Li and Gushta, 2026; Han and Rijmen, 2026), each item pair was presented twice, once in each ordering (AB and BA). A pair was classified as order-consistent when the same item was selected as more difficult under both presentations. Pilot analyses also evaluated an additional tie-breaking prompt for pairs with inconsistent AB/BA judgments. With Claude Opus, these tie-breaking judgments were only approximately 50% accurate, so the procedure was not retained for the full analysis. Pairwise judgments were instead aggregated using Bradley–Terry models (Bradley and Terry, 1952) and Davidson models (Davidson, 1970) estimated via maximum likelihood. The Bradley–Terry analysis removed all order-inconsistent comparisons, whereas the Davidson model retained them as ties. Both models produced latent scores designed to represent relative item difficulty (ACER, 2022; Han and Rijmen, 2026).

Generalization to new items was evaluated using five-fold item-stratified and passage-disjoint cross-validation. The item-stratified analysis held out individual items while maintaining approximately comparable item characteristics across folds. This was designed to mimic a scenario in which a passage is already in operational use, but new associated items have been developed. The passage-disjoint analysis held out entire connected item–passage groups. The passage-disjoint analysis reflects the scenario in which new passages and their associated items require evaluation. Within each fold, the pairwise model was first estimated using comparisons between training items. Each held-out item was then located on the latent difficulty scale using only comparisons between that item and training items, trying to identify its location on the established scale. Because the pairwise-derived scores were on an arbitrary latent scale, a linear

transformation to the empirical difficulty scale was estimated using training items only and then applied to the held-out items. Thus, the empirical item difficulty of a held-out item was not used to estimate its pairwise location or to align the latent scale.

Performance was then evaluated using measures reflecting different aspects of recovery. Pairwise accuracy indicated whether the LLM selected the item with higher empirical difficulty, while order consistency indicated whether the judgment was invariant to presentation order. Pearson correlation measured linear correspondence between the pairwise-derived and empirical difficulty, while Spearman rank correlation measured recovery of the empirical item ordering. Root mean squared error (RMSE) quantified cross-validated prediction error after alignment to the empirical difficulty scale. These measures were considered jointly rather than imposing a single threshold. Finally, potential difficulty drivers were examined by forming standardized signed differences in the metadata and LLM-generated features for each item pair. For each item pair, the response indicated which item was more difficult, with the metadata and LLM-generated features included as explanatory variables using the difference in each characteristic between the two items. The resulting slope estimates were used to identify characteristics associated with greater relative item difficulty. Separate models were fit using the LLM pairwise judgments and the empirical difficulty ordering as the response, allowing the characteristics associated with LLM-perceived difficulty to be compared with those associated with empirical item difficulty.

5 Results

Pairwise performance differed substantially across LLMs and prompting conditions, shown in Table 1. Claude Opus produced more order-consistent judgments and more accurate decisions than Qwen3-32B under both prompting conditions. This suggests that model choice has a significant impact on the overall performance of the pairwise comparison approach. The main difference between the inclusion or exclusion of item metadata and LLM-generated features was the order-consistent judgments. While there was little improvement with Claude Opus, there was a much larger improvement for Qwen3-32B. Greater order consistency, however, did not necessarily correspond to greater

accuracy among those judgments. Conditional accuracy was relatively similar across the four conditions, with Claude Opus showing slightly better performance with the exclusion of the metadata and features. However, due to the higher rate of order-consistent judgments, this led to an overall increase in the number of accurate judgments. Thus, considering conditional accuracy alone would obscure the benefit of producing a greater number of consistent, correct judgments, marked by the inclusion of the item metadata and LLM-generated features.

The models also exhibited different positional selection patterns across all of the pairwise comparisons. Claude Opus showed little positional preference under Item-Only prompting, selecting the first- and second-presented items at nearly equal rates. Under Item+Both prompting, there was a little more preference for the second-presented item. In contrast, Qwen3-32B selected the second-presented item substantially more often under both prompting conditions, although this preference was reduced when metadata and LLM-generated features were provided. These patterns are consistent with previous findings that LLM-based comparative judgments can exhibit positional bias (Li and Gushta, 2026; Han and Rijmen, 2026), while also suggesting that the magnitude of positional preference may depend on both the LLM and the information provided in the prompt. Because presentation position was not independently randomized across comparisons, these results are interpreted as descriptive positional selection patterns rather than causal estimates of presentation-order effect.

Accuracy also depended strongly on the empirical difference in difficulty between the two items, as illustrated in Table 3. Among order-consistent comparisons, judgments were only approximately 53% accurate when the items were very similar in terms of empirical difficulty. Accuracy increased steadily as the empirical difference increased reaching approximately 83–87%. Within each difficulty range, differences among LLMs and prompting conditions were comparatively small. Under the approaches evaluated in this study, pairwise judgments were therefore more successful at distinguishing items with larger empirical differences in difficulty than items with similar difficulty.

Beyond the accuracy of individual pairwise comparisons, the judgments were evaluated based on how well they recovered the empirical ordering of the items (Table 4). When judgments from both presentation orders were aggregated, Spearman rank

correlations with empirical item-difficulty orderings ranged from 0.445 to 0.541. As before, Claude Opus tended to outperform Qwen3-32B. However, the addition of item metadata and LLM-generated features improved rank recovery for Qwen3-32B, whereas no improvement was observed for Claude Opus. Focusing solely on order-consistent comparisons also produced little change, with little difference between the both-orders and order-consistent-only approaches across all four conditions. This indicates that using both presentation orders was useful for addressing positional instability, but subsequently disregarding order-inconsistent pairs did not meaningfully improve recovery of the empirical ranking. Furthermore, approximately 70% of items were placed in the correct upper or lower half of the empirical difficulty distribution and approximately 55% were placed within 20 percentile points of their empirical position. The pairwise judgments therefore recovered meaningful information about relative item difficulty, although substantial ranking error remained.

Because evaluating each pair in both presentation orders doubles the number of LLM judgments required, the AB-only and BA-only results provide an additional comparison for considering whether both presentations are necessary. For Claude Opus, rank recovery was similar whether only the AB judgments, only the BA judgments, or both orders were used. Qwen3-32B was more sensitive to this choice, particularly under Item-Only prompting, and combining both orders produced better rank recovery than either single-order approach. The rankings obtained separately from AB and BA judgments were also highly similar for Claude Opus ($\rho = 0.983$ for Item-Only and $\rho = 0.971$ for Item+Both), compared with substantially weaker agreement for Qwen3-32B ($\rho = 0.353$ and 0.741 , respectively). These comparisons do not provide a direct estimate of performance under a study designed to use only one presentation per pair because items may appear disproportionately in the first or second position within a single-order subset. Nevertheless, the results suggest that the benefit of doubling the comparisons to address presentation order may depend on the LLM, with a clearer benefit observed for Qwen3-32B than for Claude Opus in this study.

The pairwise judgments were next aggregated into latent item-difficulty scores using the Bradley-Terry and Davidson models. Across the four LLM and prompting conditions, both approaches pro-

duced moderate correspondence with empirical item difficulty. Rank recovery was nearly identical between the two scaling approaches within each condition, with Spearman correlations differing minimally. The Davidson model, however, generally produced stronger Pearson correlations with empirical difficulty than the Bradley-Terry model. This suggests that retaining order-inconsistent comparisons as ties affected the relative spacing of the estimated item scores more than the resulting ordering of the items. Consistent with the preceding analyses, Claude Opus under Item-Only prompting produced the strongest overall rank recovery, while the addition of metadata and LLM-generated features improved performance for Qwen3-32B but not for Claude Opus. As a reference, an oracle Bradley-Terry model based on perfect pairwise decisions provided an interpretable baseline for the Pearson correlation, since the pairwise-derived difficulty scores were not necessarily on the same scale as the empirical item difficulties. Since the Pearson correlation between the Oracle Bradley-Terry and the empirical difficulty was 0.961, this reveals that even perfect pairwise decisions do not completely reproduce the exact spacing of the empirical difficulty scale.

The pairwise-derived scores were also evaluated for their ability to generalize to items not used to estimate the underlying latent scale. Results from the item-stratified and passage-disjoint cross-validation analyses are shown in Tables 6 and 7, respectively. Across conditions, Davidson models generally produced stronger Pearson correlations and lower RMSE than the corresponding Bradley-Terry models, whereas differences in Spearman rank correlations were comparatively small. This suggests that retaining order-inconsistent comparisons as ties provided useful information about the relative spacing of items without substantially changing their ordering. Claude Opus under Item-Only prompting generally produced the strongest cross-validated performance. As in the preceding analyses, providing metadata and LLM-generated features did not improve recovery for Claude Opus, but generally improved performance for Qwen3-32B. This suggests that the effect of including additional information, such as metadata and LLM-generated features, on the pairwise comparisons appears to be model dependent. Future research may seek to optimize the provided metadata and/or LLM-generated features to provide more stable results across models.

Performance generally decreased under passage-disjoint cross-validation, which provided the more stringent test of generalization to items associated with previously unseen passages. However, the decline was relatively small for several of the better-performing Davidson models. For example, Claude Opus under Item-Only prompting produced Pearson correlations of 0.500 and 0.495 and RMSE values of 0.812 and 0.815 under item-stratified and passage-disjoint strategies, respectively. Thus, the recovered difficulty information was not entirely dependent on comparisons involving other items associated with the same passages. An oracle analysis further showed that perfect pairwise decisions produced essentially perfect rank recovery ($\rho \approx 1.00$), but Pearson correlations of approximately 0.96 and RMSE values of approximately 0.26 after alignment. Consequently, even perfect recovery of the pairwise ordering does not fully reproduce the exact spacing of the empirical difficulty scale, providing a reference for interpreting the scale-dependent metrics.

Finally, the difficulty-driver analyses examined whether the item characteristics associated with LLM-perceived difficulty corresponded with those associated with empirical difficulty. Several characteristics showed similar directional relationships across the empirical and LLM models (Figure 2). *Embedded inference* and *question complexity* were generally associated with greater difficulty under both definitions, while *number of cognitive steps* was consistently treated as a difficulty signal by both LLMs. However, the LLMs placed substantially greater emphasis on *number of cognitive steps* than was supported by its empirical relationship with difficulty. Thus, some characteristics appeared to represent shared difficulty signals, even when their relative importance differed between LLM judgments and student performance.

Other characteristics revealed more substantial differences between LLM-perceived and empirical difficulty. *Question explicitness/abstraction* was negatively associated with empirical difficulty ($\beta = -0.295$, $p = 0.017$) but positively associated with perceived difficulty across all four LLM conditions. *Shared wording/text echoing* showed the opposite pattern: it was positively associated with empirical difficulty ($\beta = 0.230$, $p = 0.027$) but negatively associated with LLM-perceived difficulty. *Passage-based distractors* were also positively associated with item difficulty ($\beta = 0.237$, $p = 0.027$) while contributing comparatively little

to the LLM judgments. These discrepancies indicate that the characteristics perceived by an LLM as signals of difficulty do not necessarily operate similarly for students.

More broadly, correspondence between the complete empirical and LLM difficulty-driver profiles was limited, as shown in Table 8. Depending on the LLM and prompt, only 43.8–59.4% of coefficients agreed in direction with the empirical model, and rank-order correspondence among coefficients was near zero. Nevertheless, *question complexity* and *embedded inference* were among the LLM-generated features that emerged prominently across the empirical and LLM pairwise judgments. *Number of cognitive steps* was particularly prominent in the LLM judgments despite a weaker empirical relationship. Providing metadata and LLM-generated features did not systematically improve this correspondence. The difficulty-driver analyses therefore provided evidence about the characteristics LLMs use when judging relative difficulty, while also demonstrating the importance of validating those characteristics against observed student performance.

6 Contributions

These findings suggest that LLM-based pairwise comparisons are not currently a replacement for traditional field testing with student response data, but may provide useful preliminary information about relative difficulty before such data are available. Importantly, this study evaluated a particularly fine-grained item-difficulty modeling problem: all items came from a single grade level and content area. Studies that combine items across grade levels may include broader differences in item difficulty, potentially making some relative distinctions easier to identify. In contrast, many of the Grade 6 ELA items examined here were relatively similar in empirical difficulty, and pairwise judgments approached chance-level accuracy when empirical differences were small. Accuracy increased steadily as the difference between items increased. LLM-based pairwise comparisons may therefore be more useful for identifying broader distinctions between easier and harder items than for establishing fine-grained difficulty orderings within a relatively homogeneous item bank. Such information could nevertheless support early item review, preliminary difficulty targeting, and decisions about which items warrant additional scrutiny

before field testing.

The results also demonstrate that the quality of pairwise judgments depends on both the LLM and how information is presented. Providing item metadata and LLM-generated features improved the stability and accuracy of judgments in some conditions, particularly for Qwen3-32B, but did not uniformly improve recovery of the empirical difficulty ranking. Claude Opus generally outperformed Qwen3-32B, while the models also exhibited substantially different positional selection patterns. Evaluating each pair in both presentation orders provided a useful diagnostic of this instability, but order consistency should not itself be interpreted as evidence of correctness. Moreover, the comparison of single- and both-order rankings suggests that the practical value of doubling the number of judgments may depend on the LLM being used. For assessment programs considering LLM-assisted workflows, pairwise judgments should therefore be evaluated using multiple forms of evidence, including positional stability, agreement with empirical difficulty when available, and generalization to held-out items and passages.

Finally, the framework provides information beyond a predicted item ordering. By examining characteristics associated with LLM pairwise judgments and comparing them with characteristics associated with empirical differences in difficulty, the approach can identify where LLM perceptions of difficulty align with, or diverge from, student performance. This distinction is important because characteristics that appear salient to an LLM need not have the same relationship with difficulty for students. Such information may be useful to item writers and assessment specialists seeking to understand why an item may be perceived as relatively difficult and to identify characteristics warranting attention during item development and review. Taken together, these findings position LLM-based pairwise comparison as a potential source of interpretable, supplemental pre-calibration evidence rather than a replacement for empirical calibration.

References

- ACER. 2022. Pairwise comparison method: Concept note. Technical Report WG/GAML/12, Australian Council for Educational Research.
- Peter Baldwin, Victoria Yaneva, Janet Mee, Brian E. Clauser, and Le An Ha. 2021. [Using natural language processing to predict item response times and improve test construction](#). *Journal of Educational Measurement*, 58(1):4–30.
- Ralph Allan Bradley and Milton E. Terry. 1952. [Rank analysis of incomplete block designs: I. the method of paired comparisons](#). *Biometrika*, 39(3/4):324–345.
- Virginia Braun and Victoria Clarke. 2006. [Using thematic analysis in psychology](#). *Qualitative Research in Psychology*, 3(2):77–101.
- Virginia Braun and Victoria Clarke. 2021. [One size fits all? what counts as quality practice in \(reflexive\) thematic analysis?](#) *Qualitative Research in Psychology*, 18(3):328–352.
- Roger R. Davidson. 1970. [On extending the bradley–terry model to accommodate ties in paired comparison experiments](#). *Journal of the American Statistical Association*, 65(329):317–328.
- Suhwa Han and Frank Rijmen. 2026. Scaling item difficulty using LLM-based pairwise judgments: An application of the bradley–terry–luce model. In *Annual Meeting of the National Council on Measurement in Education*, Los Angeles, CA.
- Suhwa Han, Frank Rijmen, Allison Ames Boykin, and Susan Lottridge. 2025. [Leveraging fine-tuned large language models in item parameter prediction](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*, pages 250–264, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Lanrong Li and Matthew Gushta. 2026. Using AI judges’ paired comparisons to estimate item difficulty. In *Annual Meeting of the National Council on Measurement in Education*, Los Angeles, CA.
- Eiji Muraki. 1992. [A generalized partial credit model: Application of an em algorithm](#). *Applied Psychological Measurement*, 16(2):159–176.
- Sydney Peters, Nan Zhang, Hong Jiao, Ming Li, and Tianyi Zhou. 2025. [Review of text-based approaches to item difficulty modeling in large-scale assessments](#). In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 37–47, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Alastair Pollitt. 2012. [Comparative judgement for assessment](#). *International Journal of Technology and Design Education*, 22:157–170.
- Pooya Razavi and Sonya Powers. 2026. [Estimating item difficulty using large language models and tree-based machine learning algorithms](#). *International Journal of Artificial Intelligence in Education*, pages 1–73.
- Peter Tran, Wanyong Feng, Stephen G. Sireci, Hunter McNichols, and Andrew Lan. 2025. [Using item features to calibrate educational test items: Comparing artificial intelligence and classical approaches](#). *American Journal of Education and Learning*, 10(2):178–189.
- Esther Ulitzsch, Dmitry Belov, Oliver Lüdtke, and Alexander Robitzsch. 2026. [Using item parameter predictions for reducing calibration sample requirements—a case study based on a high-stakes admission test](#). *Journal of Educational Measurement*, 63(1):1–52.

LLM	Prompt	AB Acc.	BA Acc.	Order Cons.	Cons. Acc.	Overall Correct
Claude Opus	Item-Only	67.2%	67.9%	83.7%	71.0%	59.4%
Claude Opus	Item+Both	66.7%	67.3%	86.5%	69.7%	60.3%
Qwen3-32B	Item-Only	58.8%	60.7%	53.1%	68.4%	36.3%
Qwen3-32B	Item+Both	63.5%	64.9%	69.2%	70.5%	48.8%

Table 1: Pairwise Judgment Performance Across LLMs and Prompting Conditions

LLM	Prompt	First Item	Second Item
Claude Opus	Item-Only	49.70%	50.30%
Claude Opus	Item+Both	45.37%	54.63%
Qwen3-32B	Item-Only	30.19%	69.81%
Qwen3-32B	Item+Both	35.40%	64.60%

Table 2: Positional selection patterns across LLMs and prompting conditions.

$ \Delta B $	Claude Opus		Qwen3-32B	
	Item-Only	Item+Both	Item-Only	Item+Both
≤ 0.25	52.8% (7,129)	52.6% (7,467)	53.1% (4,583)	53.4% (5,795)
0.25–0.50	60.4% (6,960)	58.9% (7,273)	58.9% (4,378)	60.4% (5,683)
0.50–1.00	69.8% (12,056)	68.5% (12,524)	67.1% (7,659)	69.2% (9,974)
1.00–2.00	79.5% (14,298)	78.2% (14,644)	75.9% (9,054)	78.7% (12,000)
> 2.00	87.4% (5,834)	85.9% (5,909)	82.7% (3,673)	85.4% (4,784)

Table 3: Accuracy and number of order-consistent judgments by empirical difficulty separation.

Ranking Approach	Claude Opus		Qwen3-32B	
	Item-Only	Item+Both	Item-Only	Item+Both
AB-Only	0.527	0.502	0.343	0.433
BA-Only	0.547	0.519	0.391	0.472
Both Orders	0.541	0.513	0.445	0.483
Order-Consistent Only	0.539	0.509	0.447	0.480

Table 4: Spearman rank correlations between pairwise-derived and empirical item-difficulty rankings.

LLM	Prompt	Model	Pearson r	Spearman ρ
Claude Opus	Item-Only	Bradley–Terry	0.316	0.539
Claude Opus	Item-Only	Davidson	0.514	0.540
Claude Opus	Item+Both	Bradley–Terry	0.267	0.517
Claude Opus	Item+Both	Davidson	0.311	0.513
Qwen3-32B	Item-Only	Bradley–Terry	0.232	0.453
Qwen3-32B	Item-Only	Davidson	0.408	0.445
Qwen3-32B	Item+Both	Bradley–Terry	0.296	0.483
Qwen3-32B	Item+Both	Davidson	0.438	0.483
Oracle Reference	—	Bradley–Terry	0.961	1.000

Table 5: Correspondence between pairwise-derived latent difficulty scores and empirical item difficulty.

LLM	Prompt	Model	Pearson r	Spearman ρ	RMSE
Claude Opus	Item-Only	Bradley–Terry	0.448	0.526	0.877
		Davidson	0.500	0.538	0.812
Claude Opus	Item+Both	Bradley–Terry	0.420	0.507	0.881
		Davidson	0.454	0.511	0.838
Qwen3-32B	Item-Only	Bradley–Terry	0.343	0.443	0.891
		Davidson	0.403	0.442	0.858
Qwen3-32B	Item+Both	Bradley–Terry	0.397	0.482	0.885
		Davidson	0.434	0.480	0.845
Oracle Reference		Bradley–Terry	0.960	1.000	0.263

Table 6: Item-stratified cross-validation performance.

LLM	Prompt	Model	Pearson r	Spearman ρ	RMSE
Claude Opus	Item-Only	Bradley–Terry	0.399	0.488	0.882
		Davidson	0.495	0.530	0.815
Claude Opus	Item+Both	Bradley–Terry	0.397	0.472	0.884
		Davidson	0.442	0.496	0.843
Qwen3-32B	Item-Only	Bradley–Terry	0.271	0.407	0.903
		Davidson	0.386	0.431	0.866
Qwen3-32B	Item+Both	Bradley–Terry	0.338	0.418	0.896
		Davidson	0.424	0.473	0.850
Oracle Reference		Bradley–Terry	0.959	1.000	0.265

Table 7: Passage-disjoint cross-validation performance.

Measure	Claude Opus		Qwen3-32B	
	Item-Only	Item+Both	Item-Only	Item+Both
Direction Agreement	59.4%	56.3%	43.8%	53.1%
Pearson Correspondence	0.163	-0.078	0.448	0.413
Spearman Correspondence	0.080	-0.034	0.114	-0.035
Top-10 Overlap	4	2	5	4

Table 8: Correspondence between LLM and empirical pairwise difficulty-driver coefficients.

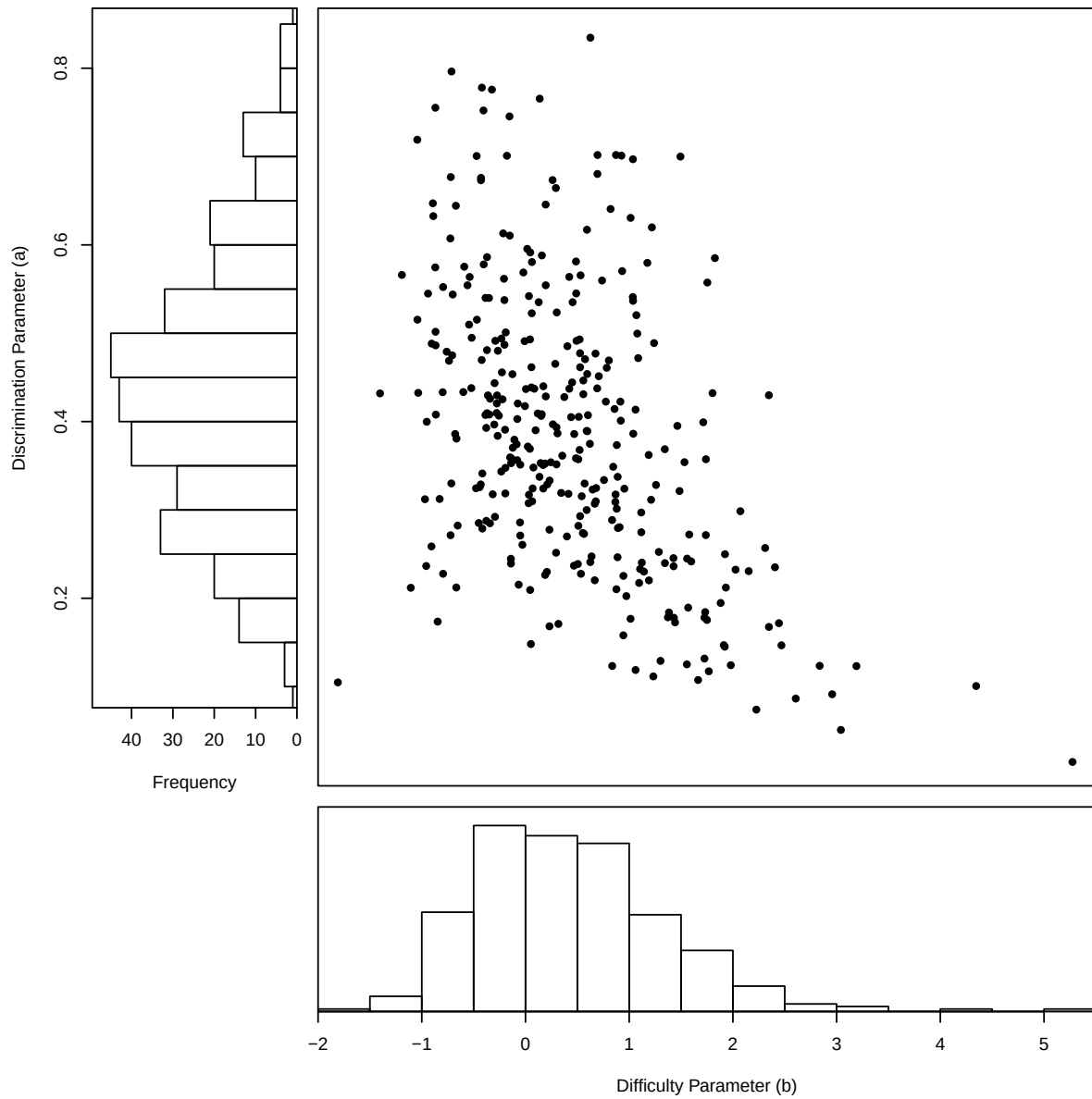


Figure 1: The distributions of empirical item difficulties and discriminations.

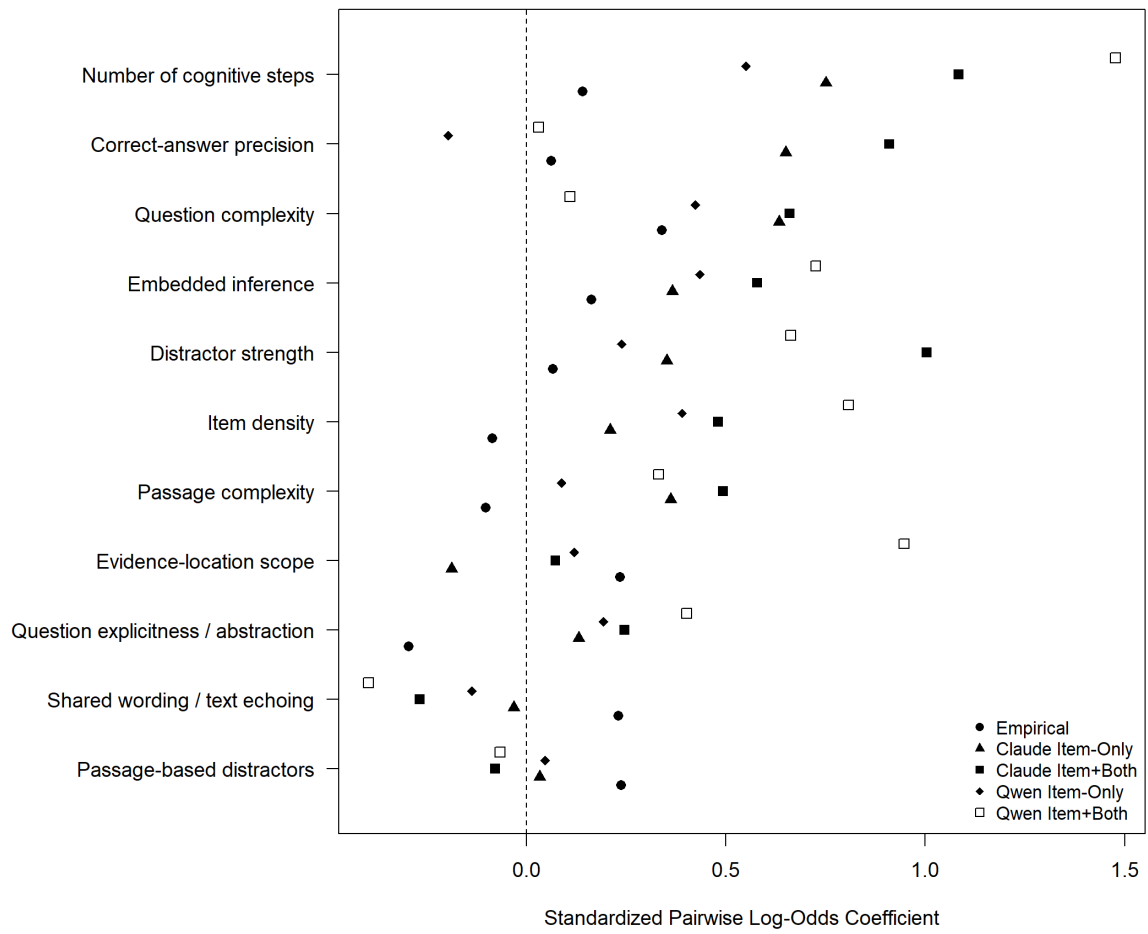


Figure 2: Selected empirical and LLM difficulty-driver coefficients.