

# Estimating LLM Grading Ability and Response Difficulty in Automatic Short Answer Grading via Item Response Theory

Longwei Cong<sup>1</sup>, Sonja Hahn<sup>1</sup>, Sebastian Gombert<sup>1</sup>, Leon Camus<sup>1</sup>,  
Hendrik Drachslér<sup>1,2</sup>, Ulf Kroehne<sup>1,3</sup>

<sup>1</sup>DIPF | Leibniz Institute for Research and Information in Education

<sup>2</sup>Faculty of Computer Science, Goethe University Frankfurt

<sup>3</sup>Chemnitz University of Technology

{l.cong, s.hahn, s.gombert, l.camus, h.drachslér, u.kroehne}@dipf.de

## Abstract

Automated short answer grading (ASAG) with large language models (LLMs) is commonly evaluated with aggregate metrics such as macro-F1 and Cohen’s kappa. However, these metrics provide limited insight into how grading performance varies across student responses of differing grading difficulty. We introduce an evaluation framework for LLM-based ASAG based on item response theory (IRT), which models grading correctness as a function of latent grader ability and response grading difficulty. This formulation enables response-level analysis of where LLM graders succeed or fail and reveals robustness differences that are not visible from aggregate scores alone. We apply the framework to 17 open-weight LLMs on the SciEntsBank and Beetle benchmarks. The results show that even models with similar overall performance differ substantially in how sharply their grading accuracy declines as response difficulty increases. In addition, confusion patterns show that errors on difficult responses concentrate disproportionately on the `partially_correct_incomplete` label, indicating a tendency toward intermediate-label collapse under ambiguity. To characterize difficult responses, we further analyze semantic and linguistic correlates of estimated difficulty. Across both datasets, higher difficulty is associated with weaker semantic alignment to the reference answer, stronger contradiction signals, and greater semantic isolation in embedding space. Overall, these results show that item response theory offers a useful framework for evaluating LLM-based ASAG beyond aggregate performance measures.

## 1 Introduction

Automatic short answer grading (ASAG) is the task of using computational methods to assess short free-text student responses to content questions (Burrows et al., 2015). Earlier ASAG approaches relied on handcrafted features and classical machine learning models (Sultan et al., 2016;

Mohler et al., 2011), followed by neural architectures and encoder-based pre-trained language models (PLMs) such as BERT (Gombert et al., 2023; Camus and Filighera, 2020). More recently, generative large language models (LLMs) have attracted increasing attention for ASAG because of their strong language understanding and multi-step reasoning capabilities (Cong et al., 2026b; Frohn et al., 2025; Ferreira Mello et al., 2025).

Despite this progress, evaluation in ASAG remains dominated by aggregate metrics such as accuracy, macro-F1, and agreement coefficients (Bonthu et al., 2021). While useful for summarizing overall system performance, these measures provide only a coarse view of grading behavior and reveal little about how performance varies across responses of differing grading difficulty (Rodriguez et al., 2021; Choi et al., 2026). This limitation is particularly important in ASAG, where student answers are often brief, diverse, and semantically ambiguous (Burrows et al., 2015). Consequently, two models with similar overall scores may nevertheless differ substantially in how well they handle genuinely difficult responses.

This motivates a more measurement-oriented perspective on LLM-based ASAG. In particular, a useful evaluation framework should help characterize which student responses are difficult to grade and whether LLM graders with similar average performance are equally robust across the difficulty spectrum. Existing ASAG research has paid relatively little attention to these questions at the response level (Haller et al., 2022).

To address this gap, we draw on item response theory (IRT), a psychometric framework that models observed outcomes as a function of latent ability and item difficulty. Prior work has shown that IRT can support more informative evaluation in NLP when benchmark instances vary substantially in difficulty (Lalor et al., 2016, 2018). We argue that IRT is well suited to ASAG because it separates latent

grader ability from response difficulty, enabling a more fine-grained analysis of grading behavior than aggregate metrics alone.

In this work, we apply an IRT-based framework to LLM-based ASAG to study grading performance at the level of individual student responses. Specifically, we make the following contributions:

- We introduce an IRT-based perspective for evaluating LLM-based ASAG, estimating latent grader abilities and response grading difficulties across benchmark datasets and multiple LLM graders.
- We analyze how grading performance changes across response difficulty levels and show that LLM graders differ substantially in their robustness as difficulty increases.
- We examine semantic and linguistic correlates of response difficulty and find that difficult responses are consistently associated with weaker semantic alignment to the reference, stronger contradiction-related signals, and greater embedding-space isolation.
- We provide error analyses showing that, on high-difficulty responses, misclassifications are not randomly distributed but increasingly concentrate in the `partially_correct_incomplete` category.

## 2 Background

### 2.1 Evaluation Methods in LLM-based ASAG

Evaluation in ASAG has traditionally focused on benchmark performance over labeled response sets, typically reported using aggregate metrics such as accuracy, F1, Cohen’s kappa, RMSE, and correlation-based measures (Haller et al., 2022; Burrows et al., 2015; Sultan et al., 2016). Recent LLM-based ASAG studies largely follow this paradigm, assessing LLM graders through overall label agreement or classification performance (Ferreira Mello et al., 2025; Frohn et al., 2025; Cong et al., 2026b).

Yet recent research on LLM evaluation suggests that aggregate benchmark scores alone may provide only a limited view of model capability (Zhou et al., 2025; Li et al., 2025). Prior work shows that benchmark-level averages can obscure substantial variation in item properties such as difficulty and discriminability, while offering limited separability

among strong models (Lalor et al., 2016, 2018). Similarly, recent work on LLM-as-a-Judge argues that LLM-based evaluators should themselves be evaluated systematically, with reliability, bias, and robustness as central concerns (Gu et al., 2025).

Within ASAG, however, the response-level determinants of grading difficulty remain underexplored, particularly for LLM-based grading. Understanding which properties of student answers make them difficult to grade is important both for interpretability and for deployment, especially in human-in-the-loop settings where difficulty estimates could guide targeted review and improve grading efficiency. Although recent work has examined confidence and uncertainty in ASAG (Funayama et al., 2022; Bexte et al., 2024; Cong et al., 2026a), these signals are typically model-derived and thus only indirectly informative about the characteristics of the response itself. Analyzing difficulty as a property of the student response instead offers a more interpretable and educationally grounded perspective on the failure modes of LLM-based grading.

### 2.2 Item Response Theory

Item Response Theory (IRT) provides a principled framework by disentangling latent responder ability from item properties (van der Linden, 2016). In the simplest Rasch formulation, the probability that responder  $m$  correctly answers item  $n$  is modeled as

$$P(y_{mn} = 1 | \theta_m, b_n) = \sigma(\theta_m - b_n),$$

where  $\theta_m$  denotes the latent ability of responder  $m$ ,  $b_n$  the difficulty of item  $n$ , and  $\sigma(\cdot)$  the logistic sigmoid. More flexible variants, such as 2PL models, additionally model item discrimination, allowing some items to be more informative than others (Hambleton and Swaminathan, 2013).

Recent work has begun to apply IRT to LLM evaluation in several ways, including diagnosing the reliability of LLM-as-a-judge systems (Choi et al., 2026) and modeling latent LLM abilities together with query difficulty and discrimination for routing and performance prediction (Chen et al., 2025). In educational technology, related work has also shown that IRT can be integrated with LLM-based modeling to predict or align item difficulty in assessment settings (Scarlatos et al., 2025). Collectively, these developments suggest that psychometric modeling offers a robust lens for analyzing LLM behavior through a measurement-grounded perspective.

### 3 Method

#### 3.1 Datasets and LLMs

We evaluate our approach on the five-way SciEnts-Bank and Beetle datasets (Dzikovska et al., 2013), two widely used benchmarks for ASAG. Released as part of the SemEval student-response analysis task, both datasets contain science-domain questions, reference answers, and student responses annotated by experts. An important property of these benchmarks is their fine-grained five-way label space, which enables analysis beyond binary correctness. Each instance consists of a question, a single reference answer, a student response, and a gold label from the set `correct`, `contradictory`, `partially_correct_incomplete`, `irrelevant`, and `non_domain`. An important difference between the two datasets is that Beetle is drawn from a relatively narrow tutorial domain in basic electricity and electronics, whereas SciEntsBank covers assessment responses across a substantially broader set of science topics (Dzikovska et al., 2013).

As LLM graders, we select 17 open-weight models covering multiple model families including Google, Qwen, Microsoft, Meta-Llama, MistralAI, and OpenChat, with parameter sizes ranging from 0.8B to 14B. The complete list of models is provided in Table 4 in Appendix A.2. The goal is to capture substantial diversity in grading behavior, including variation associated with model size, architectural lineage, and training characteristics.

#### 3.2 IRT Estimation of Grader Ability and Response Difficulty

We treat each LLM as a grader and each student response as an item to be evaluated. Given a question, a reference answer, and a student response, each model is prompted to assign exactly one label from the five benchmark categories. Decoding is performed greedily with temperature set to 0 to ensure deterministic and reproducible model outputs.

Let  $i \in \{1, \dots, M\}$  index LLM graders and  $j \in \{1, \dots, J\}$  index student-response instances. For each grader–response pair, we define a binary observation

$$y_{ij} = \begin{cases} 1 & \text{if grader } i \text{ is correct on response } j, \\ 0 & \text{otherwise.} \end{cases}$$

Because multiple student responses are associated with the same question, we include grader-specific testlet effects to capture residual question-specific variation beyond overall grader ability and

response difficulty. Specifically, we extend the Rasch model with testlet effects and fit a testlet response theory (Wainer et al., 2007) variant:

$$P(y_{ij} = 1) = \sigma(\theta_i - b_j + u_{i,t(j)}),$$

where  $\theta_i$  denotes the latent grading ability of grader  $i$ ,  $b_j$  denotes the grading difficulty of student-response instance  $j$ , and  $t(j)$  denotes the question-level testlet to which response instance  $j$  belongs. The term  $u_{i,t(j)}$  is a grader-specific deviation associated with testlet  $t(j)$ , allowing graders to exhibit question-specific performance offsets beyond their overall ability. We use a 1PL testlet formulation rather than a 2PL variant because a 2PL model would introduce an additional discrimination parameter for each response instance. Given the large number of response instances, we therefore adopt the 1PL formulation and account for question-level dependence through testlet deviations.

Parameters are estimated by minimizing the regularized negative log-likelihood:

$$\mathcal{L} = - \sum_{i,j} \left[ y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij}) \right] + \frac{\lambda_\theta}{2} \|\theta\|_2^2 + \frac{\lambda_b}{2} \|b\|_2^2 + \frac{\lambda_u}{2} \|u\|_2^2, \quad (1)$$

Following our implementation, we set  $\lambda_\theta = \lambda_b = 1.0$  and  $\lambda_u = 5.0$ . To ensure identifiability,  $\theta$  and  $b$  are mean-centered within the fitted dataset, and the testlet effects are centered across graders within each testlet. Optimization is performed using L-BFGS (Liu and Nocedal, 1989).

##### 3.2.1 Parameter Recovery

To assess whether the latent parameters are recoverable in practice, we conduct a parameter recovery study using the fitted IRT model as the data-generating process, following standard practice in IRT evaluation (Harwell, 1997). Specifically, we treat the fitted model as a generative data source and conduct a simulation-based parameter recovery analysis. For each replication, we use the estimated grader abilities, response difficulties, and grader-specific testlet effects to compute the probability of correct grading for every grader–response pair under the fitted 1PL testlet model. We then sample a new binary correctness matrix from these probabilities, re-fit the model to the simulated data using the same estimation procedure as for the original data, and compare the recovered parameters with

the original fitted parameters after alignment to a common scale.

Recovery is quantified using Pearson correlation, RMSE, and MAE, after aligning the recovered parameters to the original scale via mean and standard-deviation matching. We report these metrics for grader abilities and response difficulties. This analysis assesses whether the inferred latent quantities are sufficiently recoverable to support downstream interpretation.

### 3.2.2 Split-Half Stability

To assess empirical robustness on real data, we conduct a split-half stability analysis. Whereas parameter recovery evaluates recoverability under synthetic responses generated from the fitted model, split-half stability examines whether the inferred parameters remain consistent when the model is re-estimated on disjoint subsets of the observed data.

We use two complementary procedures. To assess grader-ability stability, we randomly split responses into two halves, fit the model separately on each half using the full set of graders, and compare the resulting ability estimates. Because our main model includes question-level testlet effects, response splits are performed within each question. To assess response-difficulty stability, we randomly split graders into two halves, fit the model separately on each subset using the full set of responses, and compare the resulting difficulty estimates.

We repeat both procedures over 10 random replications and quantify stability using Pearson and Spearman correlations, together with RMSE and MAE after mean–standard deviation alignment to account for scale indeterminacy. High split-half agreement indicates that the inferred latent quantities capture reproducible structure in the observed data.

### 3.3 Difficulty-Conditioned Analysis of LLM Grader Behavior

To analyze how grading behavior changes as response difficulty increases, we use the estimated response difficulty  $b_j$  as the primary stratification variable. We partition responses into quantile-based difficulty bins and compute each grader’s empirical accuracy within each bin. This stratified evaluation complements aggregate metrics by revealing not only whether performance declines with increasing difficulty, but also whether graders differ in their sensitivity to difficult responses.

Using these bins, we perform two analyses. First, we examine the relationship between difficulty and grading accuracy to assess how performance changes across the difficulty spectrum. Second, we inspect bin-specific confusion matrices to identify how error patterns shift with increasing difficulty.

### 3.4 Semantic and Linguistic Correlates of Difficulty

To examine which response properties are associated with grading difficulty, we extract a focused set of semantic and linguistic features for each student response, motivated by prior work in ASAG (Mohler and Mihalcea, 2009; Burrows et al., 2015; Dzikovska et al., 2013). These features include response length, lexical diversity, unigram and bigram overlap with the reference answer, semantic similarity to the reference answer, NLI-based features computed on the reference–response pair, embedding-space neighborhood features, and the number of missing reference segments. Sentence-embedding features are computed using all-MiniLM-L6-v2 (Reimers and Gurevych, 2019), and NLI-based features are obtained using a ModernBERT-based NLI model (Warner et al., 2025). A detailed description of the features is provided in Appendix A.1.

We treat these features as correlates of estimated response difficulty. For each dataset, we compute both Pearson and Spearman correlations between each feature and the estimated difficulty  $b_j$ , and apply the Benjamini–Hochberg procedure to control the false discovery rate under multiple comparisons (Benjamini and Hochberg, 1995).

## 4 Results

We report three sets of results. First, we assess the stability of the proposed IRT framework and its latent estimates of grader ability and response difficulty. Second, we examine how LLM grading performance degrades as difficulty increases, including changes in confusion patterns on more difficult responses. Third, we analyze how estimated difficulty relates to semantic alignment and response atypicality.

### 4.1 Parameter Recovery and Split-Half Stability

The fitted IRT model converged successfully on both datasets and produced a broad range of latent grader-ability estimates across the 17 evaluated LLMs, indicating substantial heterogeneity in

Parameter	Pearson $r$	RMSE	MAE	Conv.
<i>SciEntsBank</i>				
grader ability ( $\theta$ )	0.999	0.029	0.024	1.000
response difficulty ( $b$ )	0.896	0.505	0.404	1.000
<i>Beetle</i>				
grader ability ( $\theta$ )	0.998	0.036	0.029	1.000
response difficulty ( $b$ )	0.895	0.497	0.398	1.000

Table 1: Parameter recovery results over 10 replications on SciEntsBank and Beetle. Values are mean recovery statistics after mean–std alignment to the ground-truth scale. Conv. denotes convergence rate.

grading ability. The full model-level results are reported in Table 4 in Appendix A.2.

Parameter recovery results in Table 1 further support the stability of these latent estimates under the fitted model. For grader ability, recovery was nearly perfect on both datasets, with mean aligned Pearson correlations of 0.999 on SciEntsBank and 0.998 on Beetle, together with very low RMSE/MAE values (0.029/0.024 and 0.036/0.029, respectively). Response difficulty estimates were also recovered well, with Pearson correlations of 0.896 on SciEntsBank and 0.895 on Beetle, and comparable error levels (RMSE = 0.505 and 0.497; MAE = 0.404 and 0.398). All recovery runs converged successfully on both datasets.

Complementary split-half analyses on the observed data yielded a similar pattern (Table 2). Grader ability estimates were highly stable across response halves, whereas response difficulty estimates showed moderate-to-strong stability across grader halves, indicating that the inferred difficulty structure is reproducible but noisier than the model ability scale. Taken together, these results suggest that the IRT framework yields identifiable and empirically robust estimates of both latent grading ability and response difficulty, providing a sound measurement-oriented foundation for the subsequent analyses.

## 4.2 Differential Performance Degradation Across Difficulty Levels

LLM graders differ substantially in how their grading performance degrades as response difficulty increases. We grouped responses into five quantile-based bins according to their estimated difficulty  $b$ , from the easiest ( $B1$ ) to the hardest ( $B5$ ). Across both SciEntsBank and Beetle, accuracy declined monotonically as difficulty increased. At the same

Parameter	Pearson $r$	RMSE	MAE	Conv.
<i>SciEntsBank</i>				
grader ability ( $\theta$ )	0.996	0.051	0.041	1.000
response difficulty ( $b$ )	0.791	0.595	0.461	1.000
<i>Beetle</i>				
grader ability ( $\theta$ )	0.994	0.055	0.043	1.000
response difficulty ( $b$ )	0.786	0.595	0.464	1.000

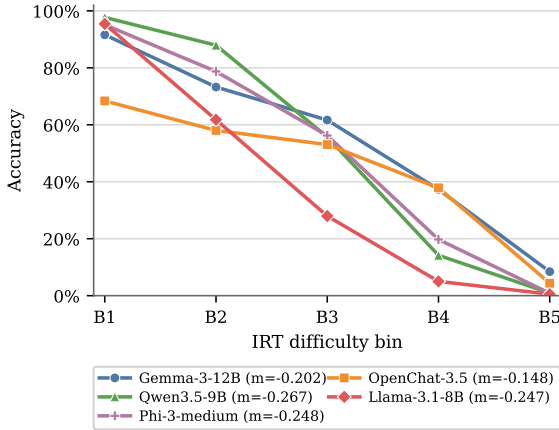
Table 2: Split-half stability results over 10 random replications on SciEntsBank and Beetle. Values are mean agreement statistics after mean–std alignment across halves. Conv. denotes convergence rate.

time, the rate of decline varied markedly across models, indicating substantial differences in robustness to difficult responses.

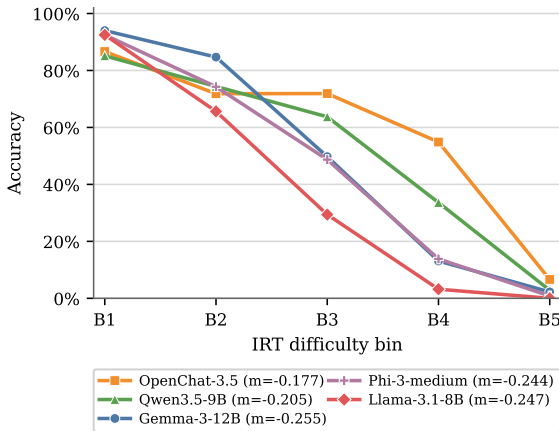
On SciEntsBank, all examined models showed sharp performance degradation toward the most difficult bins, but the extent of this decline differed. Gemma-3-12B remained the most robust overall, retaining 16.6% accuracy on  $B5$ , whereas Qwen3.5-9B, Phi-3-medium, OpenChat-3.5, and Llama-3.1-8B fell to 4.1%, 5.8%, 15.9%, and 2.1%, respectively. On Beetle, the same overall pattern held, although the relative robustness profile differed. OpenChat-3.5 degraded more gradually than the other models and retained the strongest performance in the upper-difficulty bins, reaching 67.2% on  $B4$  and 12.9% on  $B5$ . By contrast, Gemma-3-12B and Phi-3-medium dropped much more sharply after  $B3$ , while Llama-3.1-8B approached near-zero performance already by  $B4$ .

Additional results for the full set of models are provided in Fig. 3 in Appendix A.2. Notably, the difficulty–performance slopes are broadly consistent across the two datasets. The cross-dataset association is strong, with a Pearson correlation of 0.838 and a Spearman correlation of 0.706. This suggests that sensitivity to increasing response difficulty is not merely dataset-specific, but instead reflects a comparatively stable characteristic of LLM grading behavior.

To further characterize this degradation, we examined confusion patterns across difficulty bins. As shown in Fig. 2 and Fig. 4 in Appendix A.2, errors in high-difficulty responses are not randomly distributed across labels, but instead become increasingly structured. Across both datasets, correct, contradictory, irrelevant, and non\_domain responses are increasingly mapped to partially\_correct\_incomplete (PCI) as difficulty rises, indicating a progressive loss of fine-



(a) SciEntsBank



(b) Beetle

Figure 1: Model accuracy across ordered response-difficulty bins on (a) SciEntsBank and (b) Beetle. Here,  $m$  denotes the slope obtained by linearly regressing model accuracy on the order of the IRT-based difficulty bins.

grained multi-class discrimination. At the same time, difficult PCI responses are often misclassified as correct, suggesting that higher difficulty induces both collapse toward an intermediate category and overly optimistic judgments of ambiguous answers.

### 4.3 Semantic Misalignment and Response Atypicality as Correlates of Difficulty

As shown in Table 3, the correlation analysis reveals a clear and largely consistent pattern across SciEntsBank and Beetle. Overall, the strongest associations with response difficulty come from features that capture semantic alignment with the reference answer and the semantic typicality of a student response relative to other responses. In both datasets, higher difficulty is associated

with lower NLI entailment to the reference answer (SciEntsBank: Pearson  $r = -0.300$ , Spearman  $\rho = -0.303$ ; Beetle: Pearson  $r = -0.261$ , Spearman  $\rho = -0.227$ ) and a smaller entailment-contradiction margin (SciEntsBank:  $r = -0.291$ ,  $\rho = -0.306$ ; Beetle:  $r = -0.235$ ,  $\rho = -0.232$ ). Likewise, semantic similarity to the reference answer and lexical overlap with the reference answer, including both unigram and bigram overlap, are negatively correlated with difficulty in both datasets. These results indicate that responses that are semantically closer to the reference answer tend to be easier for LLM graders to evaluate correctly.

By contrast, features reflecting semantic disagreement or atypicality show positive associations with difficulty. In both datasets, the NLI contradiction score is positively correlated with response difficulty (SciEntsBank:  $r = 0.211$ ,  $\rho = 0.309$ ; Beetle:  $r = 0.136$ ,  $\rho = 0.216$ ). A similar pattern holds for embedding-space neighborhood distance. Both average and minimum  $k$ -nearest-neighbor distance are positively associated with difficulty across datasets. This suggests that responses that are more isolated in semantic space, and therefore less typical relative to the broader response distribution, are more difficult for LLM graders to grade reliably.

At the same time, not all features are equally informative. Some shallow surface-level properties, such as token count and type-token ratio, show weak or inconsistent associations across datasets. The number of missing reference segments is positively related to difficulty in SciEntsBank, but much weaker in Beetle.

## 5 Discussion

Our findings support a difficulty-aware perspective on evaluating LLM-based ASAG. Although aggregate measures such as accuracy, macro-F1, and agreement coefficients are useful summaries of overall performance, they do not reveal whether a model remains robust across responses of differing grading difficulty, a limitation also noted in prior IRT-based work on NLP evaluation (Lalor et al., 2016, 2018). The simulation-based recovery and split-half analyses further support the stability of the proposed measurement framework, suggesting that the estimated latent structure is not arbitrary and remains reasonably consistent across repeated analyses. Split-half results also show that grader ability estimates are more stable than response dif-

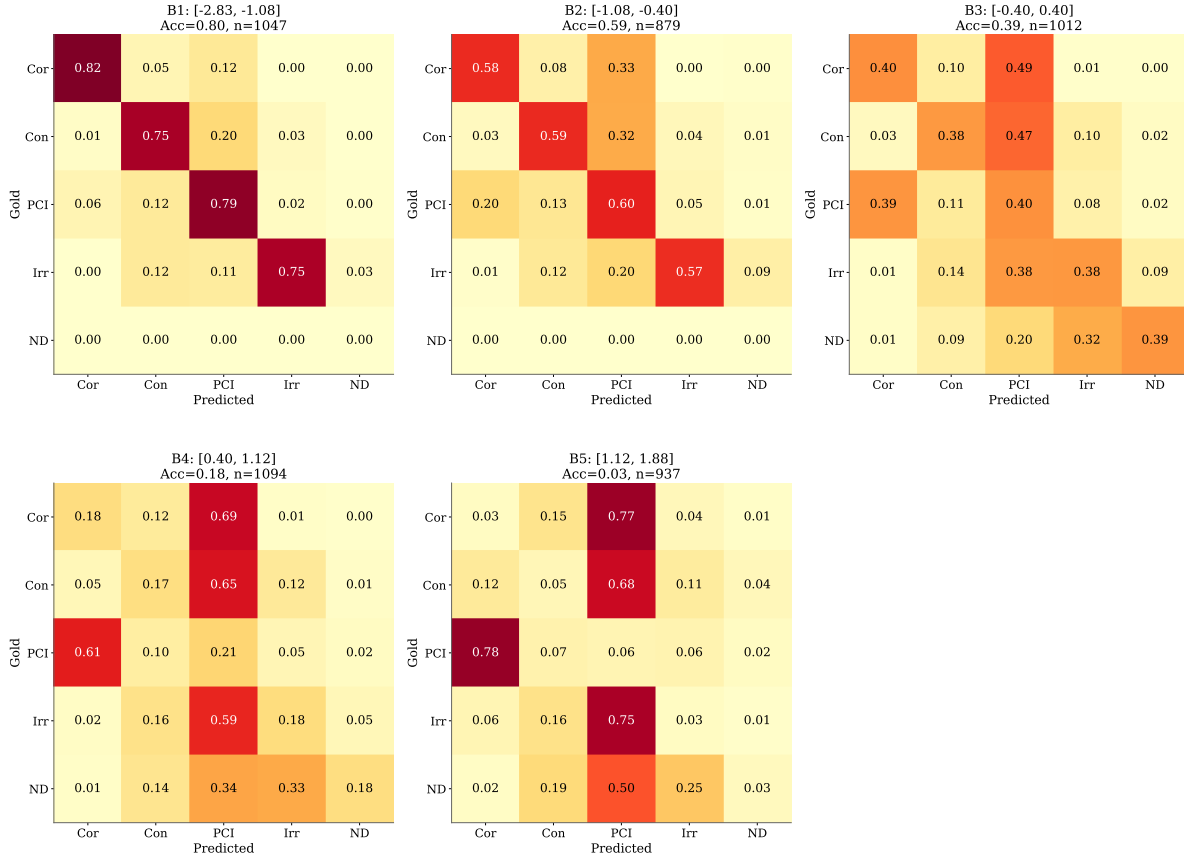


Figure 2: Confusion matrices on SciEntsBank across five bins of IRT-derived response difficulty  $b$ .  $n$  denotes the number of responses in each bin, and Acc denotes overall grading accuracy.

difficulty estimates, likely because the former are estimated from many responses, whereas the latter are supported by far fewer grader judgments.

This difficulty-aware analysis therefore provides a more informative account of where LLM graders succeed and where they fail. More specifically, the results suggest that LLM graders should be compared not only by their overall performance, but also by how robustly they handle increasingly difficult responses. Models with similar aggregate scores may nevertheless differ substantially in the rate and form of performance degradation across difficulty levels. From this perspective, a grader that performs well on easy cases but deteriorates sharply on difficult responses is qualitatively different from one whose performance declines more gradually, even when their aggregate scores are similar. From a robustness standpoint, the latter may be better suited for practical deployment.

The error analysis suggests that high-difficulty responses do not simply produce more mistakes at random. Instead, errors become increasingly concentrated in the `partially_correct_incomplete` category.

This pattern indicates that, under uncertainty, LLM graders often default to an intermediate label rather than making a sharper distinction. One plausible interpretation is that `partially_correct_incomplete` acts as a default intermediate category when evidence is weak, mixed, or only partially supportive. In such cases, the model preserves topical plausibility without committing to full correctness.

The feature analysis helps explain why such responses are difficult. Across datasets, higher difficulty is most consistently associated with weaker semantic alignment to the reference answer, stronger contradiction-related signals, and greater isolation in embedding space. These findings suggest that difficult responses are typically not arbitrary outliers, but responses whose meaning is incomplete, weakly expressed, internally mixed, or only partially compatible with the reference answer. At the same time, the correlations are moderate rather than large, which implies that difficulty is not reducible to any single response property. Instead, grading difficulty appears to emerge from the interaction of multiple semantic factors, including

Feature	SciEntsBank		Beetle	
	Pearson $r$	Spearman $\rho$	Pearson $r$	Spearman $\rho$
NLI entailment to reference	-0.300***	-0.303***	-0.261***	-0.227***
Entailment-contradiction margin	-0.291***	-0.306***	-0.235***	-0.232***
Average kNN distance	+0.206***	+0.250***	+0.210***	+0.200***
NLI contradiction to reference	+0.211***	+0.309***	+0.136***	+0.216***
Bigram overlap with reference	-0.247***	-0.222***	-0.147***	-0.110***
Semantic similarity to reference	-0.204***	-0.220***	-0.190***	-0.185***
Minimum kNN distance	+0.183***	+0.218***	+0.187***	+0.190***
Unigram overlap with reference	-0.203***	-0.192***	-0.160***	-0.127***
NLI neutrality to reference	+0.134***	+0.162***	+0.133***	+0.137***
Number of missing reference segments	+0.133***	+0.169***	+0.040	+0.027
Type-token ratio	+0.010	+0.050*	-0.055***	-0.050***
Token count	-0.008	-0.042	+0.050***	+0.064***

Table 3: Correlation between response difficulty and linguistic/semantic features on SciEntsBank and Beetle. Cell values are Pearson  $r$  and Spearman  $\rho$ ; superscripts indicate Benjamini-Hochberg adjusted significance ( $*q < .05$ ,  $**q < .01$ ,  $***q < .001$ ). Features are ordered by mean absolute Pearson correlation across the two datasets.

alignment, contradiction, completeness, and local response-space structure.

Taken together, these results have practical implications for LLM-assisted grading. Most importantly, they suggest that not all responses should be treated as equally suitable for fully automatic scoring. If difficulty can be estimated directly or approximated from response features, it could support selective grading pipelines in which straightforward cases are handled automatically while difficult cases are routed to human review. Compared with uncertainty estimates derived solely from the model itself (Funayama et al., 2022; Bexte et al., 2024; Shorinwa et al., 2025), response-based difficulty also offers a more interpretable account of why particular answers are difficult to grade. This may be especially valuable in educational settings, where reliability, interpretability, and fairness are as important as efficiency (Madnani and Cahill, 2018).

The findings also have implications for rubric design, which has recently emerged as an important direction in ASAG research (Frohn et al., 2025; Cong et al., 2026b; Gombert et al., 2026). The systematic tendency to over-predict partially\_correct\_incomplete on difficult responses suggests that current prompting setups may not provide sufficiently clear decision boundaries between neighboring categories. More explicit rubric-based prompting or decomposed decision procedures may help address this problem. For example, grading could be structured into sequential sub-decisions such as topical relevance, presence of correct propositions, contradiction to the reference, and completeness of required content. This kind

of scaffolded evaluation may reduce the model’s tendency to collapse toward an intermediate class when evidence is ambiguous.

## 6 Conclusion

Overall, the present study shows that IRT offers a useful framework for analyzing LLM graders beyond aggregate performance. Rather than asking only which model achieves the highest average score, this perspective makes it possible to ask which responses are difficult, why they are difficult, and which models remain more robust as difficulty increases. For ASAG, this shift from average performance to measurement-oriented analysis provides a more interpretable basis for both evaluation and system design.

## 7 Limitations

This study has several limitations. First, the number of LLM graders is limited. Although we evaluate 17 models across multiple families and parameter scales, the estimated latent ability scale and robustness patterns should be interpreted with respect to this particular model set. Second, the analysis is restricted to two benchmark ASAG datasets and a specific five-way label schema. Although the main patterns are consistent across SciEntsBank and Beetle, it remains unclear how well they generalize to other domains, rubrics, or grading schemes. Third, our IRT formulation binarizes grading correctness and therefore does not model the five-way label structure directly. Future work could use polytomous IRT models to capture more fine-grained differences among label categories and error types. Finally, the feature analysis is correlational. The

observed associations help characterize difficult responses, but they do not establish causal determinants of grading difficulty. Investigating these causal mechanisms remains an important direction for future work.

## Acknowledgments

This research was conducted within the project “Assessment for Learning with AI (ALwAI)” funded by the Leibniz Association under the Leibniz Competition (project no. T163/2024).

## References

- Yoav Benjamini and Yoel Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.
- Marie Bexte, Andrea Horbach, Lena Schützler, Oliver Christ, and Torsten Zesch. 2024. Scoring with confidence?—exploring high-confidence scoring for saving manual grading effort. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 119–124.
- Sridevi Bonthu, S Rama Sree, and MHM Krishna Prasad. 2021. Automated short answer grading using deep learning: A survey. In *International cross-domain conference for machine learning and knowledge extraction*, pages 61–78. Springer.
- Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25(1):60–117.
- Leon Camus and Anna Filighera. 2020. Investigating transformers for automatic short answer grading. In *International conference on artificial intelligence in education*, pages 43–48. Springer.
- Jianhao Chen, Chenxu Wang, Gengrui Zhang, Peng Ye, Lei Bai, Wei Hu, Yuzhong Qu, and Shuyue Hu. 2025. Learning compact representations of llm abilities via item response theory. *arXiv preprint arXiv:2510.00844*.
- Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. 2026. Diagnosing the reliability of llm-as-a-judge via item response theory. *arXiv preprint arXiv:2602.00521*.
- Longwei Cong, Sonja Hahn, Sebastian Gombert, Leon Camus, Hendrik Drachler, and Ulf Kroehne. 2026a. Confidence estimation in automatic short answer grading with llms. *Preprint*, arXiv:2605.00200.
- Longwei Cong, Leon Hammerla, Sonja Hahn, Sebastian Gombert, Hendrik Drachler, and Ulf Kroehne. 2026b. Automatic short answer grading with LLMs: From memorization to reasoning. In *Proceedings of the 16th International Learning Analytics and Knowledge Conference*, New York, NY, USA. Association for Computing Machinery.
- Myroslava Dzikovska, Rodney Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa Trang Dang. 2013. SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 263–274, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Rafael Ferreira Mello, Cleon Pereira Junior, Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Newarney Costa, Geber Ramalho, and Dragan Gasevic. 2025. Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th international learning analytics and knowledge conference*, pages 93–103.
- Scott Frohn, Tyler Burleigh, and Jing Chen. 2025. Automated scoring of short answer questions with large language models: Impacts of model, item, and rubric design. In *International Conference on Artificial Intelligence in Education*, pages 44–51. Springer.
- Hiroaki Funayama, Tasuku Sato, Yuichiro Matsubayashi, Tomoya Mizumoto, Jun Suzuki, and Kentaro Inui. 2022. Balancing cost and quality: an exploration of human-in-the-loop frameworks for automated short answer scoring. In *International conference on artificial intelligence in education*, pages 465–476. Springer.
- Sebastian Gombert, Daniele Di Mitri, Onur Karademir, Marcus Kubsch, Hannah Kolbe, Simon Tautz, Adrian Grimm, Isabell Bohm, Knut Neumann, and Hendrik Drachler. 2023. Coding energy knowledge in constructed responses with explainable nlp models. *Journal of Computer Assisted Learning*, 39(3):767–786.
- Sebastian Gombert, Zhifan Sun, Fabian Zehner, Jannik Lossjew, Tobias Wyrwich, Berrit Katharina Czinczel, David Bednorz, Marcus Kubsch, Daniele Di Mitri, Knut Neumann, and Hendrik Drachler. 2026. Are rubrics all you need? towards rubric-based automatic short answer scoring via guided rubric–answer alignment. In *Proceedings of the 16th International Learning Analytics and Knowledge Conference*.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

- Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. 2022. Survey on automated short answer grading with deep learning: from word embeddings to transformers. *arXiv preprint arXiv:2204.03503*.
- Ronald K Hambleton and Hariharan Swaminathan. 2013. *Item response theory: Principles and applications*. Springer Science & Business Media.
- Michael R Harwell. 1997. Analyzing the results of monte carlo studies in item response theory. *Educational and Psychological Measurement*, 57(2):266–279.
- John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. [Understanding deep learning performance through an examination of test set difficulty: A psychometric case study](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716, Brussels, Belgium. Association for Computational Linguistics.
- John P. Lalor, Hao Wu, and Hong Yu. 2016. [Building an evaluation scale using item response theory](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. [From generation to judgment: Opportunities and challenges of LLM-as-a-judge](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.
- Dong C Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528.
- Nitin Madnani and Aoife Cahill. 2018. [Automated scoring: Beyond natural language processing](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1099–1109, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. [Learning to grade short answer questions using semantic similarity measures and dependency graph alignments](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 752–762, Portland, Oregon, USA. Association for Computational Linguistics.
- Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 567–575.
- Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992.
- Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. [Evaluation examples are not equally informative: How should that change NLP leaderboards?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.
- Alexander Scarlatos, Nigel Fernandez, Christopher Ormerod, Susan Lottridge, and Andrew Lan. 2025. Smart: Simulated students aligned with item response theory for question difficulty prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25082–25105.
- Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38.
- Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. 2016. [Fast and easy short answer grading with high accuracy](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1070–1075, San Diego, California. Association for Computational Linguistics.
- Wim J. van der Linden. 2016. *Handbook of item response theory*, volume 1. CRC press Boca Raton, FL.
- Howard Wainer, Eric T Bradlow, and Xiaohui Wang. 2007. *Testlet response theory and its applications*. Cambridge University Press.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Hongli Zhou, Hui Huang, Ziqing Zhao, Lvyuan Han, Huicheng Wang, Kehai Chen, Muyun Yang, Wei Bao, Jian Dong, Bing Xu, and 1 others. 2025. Lost in benchmarks? rethinking large language model benchmarking with item response theory. *arXiv preprint arXiv:2505.15055*.

## A Appendix

### A.1 Feature Definitions

For each student response, we compute a set of lexical, semantic, and distributional features with respect to the reference answer and the question. Let  $a_i$  denote the student answer for item  $i$ ,  $r_i$  the corresponding reference answer, and  $q_i$  the question. All lexical features are computed from lower-case regex-based tokens extracted with the pattern  $[A-Za-z0-9']^+$ . Thus, throughout this appendix, “token” refers to these surface tokens rather than model-specific subword tokens.

**Lexical overlap features.** Let  $U(a_i)$  and  $U(r_i)$  denote the sets of unique unigrams in the student answer and reference answer, respectively. The unigram overlap with the reference is defined as

$$\text{UnigramOverlap}(i) = \frac{|U(a_i) \cap U(r_i)|}{|U(r_i)|}.$$

Similarly, letting  $B(a_i)$  and  $B(r_i)$  denote the sets of unique bigrams in the student answer and reference answer, the bigram overlap is

$$\text{BigramOverlap}(i) = \frac{|B(a_i) \cap B(r_i)|}{|B(r_i)|}.$$

Both measures therefore quantify reference-normalized lexical coverage rather than Jaccard similarity.

The token count is simply the number of tokens in the student answer,

$$\text{TokenCount}(i) = |a_i|,$$

and the type-token ratio is defined as

$$\text{TTR}(i) = \frac{|\text{unique tokens in } a_i|}{|a_i|}.$$

**Semantic similarity features.** We encode student answers, reference answers, and questions using Sentence-BERT embeddings. All embeddings are  $\ell_2$ -normalized, so cosine similarity reduces to a dot product. The semantic similarity to the reference is computed as

$$\text{SimRef}(i) = \cos(\mathbf{e}(a_i), \mathbf{e}(r_i)).$$

**NLI-based features.** We apply a pretrained natural language inference model to each reference–answer pair, using the reference answer as premise

and the student answer as hypothesis. This yields three probabilities:

$$P_{\text{ent}}(i), \quad P_{\text{con}}(i), \quad P_{\text{neu}}(i),$$

corresponding to entailment, contradiction, and neutrality. From these, we derive the following features:

$$\begin{aligned} \text{NLIEntail}(i) &= P_{\text{ent}}(i), \\ \text{NLIContradict}(i) &= P_{\text{con}}(i), \\ \text{NLINeutral}(i) &= P_{\text{neu}}(i), \\ \text{NLIMargin}(i) &= P_{\text{ent}}(i) - P_{\text{con}}(i). \end{aligned}$$

**Distributional neighborhood features.** To quantify how typical or atypical a student response is relative to other responses in the same dataset, we compute pairwise cosine similarities among all student-answer embeddings. For each answer, we identify its  $k$  nearest neighbors in embedding space (excluding the answer itself), with  $k = 5$  in our implementation. The average  $k$ NN distance is defined as

$$\text{AvgKNNDist}(i) = 1 - \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \cos(\mathbf{e}_i, \mathbf{e}_j),$$

where  $\mathbf{e}_i = \mathbf{e}(a_i)$  denotes the embedding of answer  $a_i$ , and  $\mathcal{N}_k(i)$  denotes the set of indices of the  $k$  most similar other student answers. The minimum  $k$ NN distance is defined as

$$\text{MinKNNDistance}(i) = 1 - \max_{j \neq i} \cos(\mathbf{e}(a_i), \mathbf{e}(a_j)).$$

Higher values indicate that a response is more isolated in the embedding space.

**Reference-segment coverage.** To approximate coverage of distinct parts of the reference answer, we split each reference answer into segments using sentence-like punctuation (. ; : ! ?), tokenize each segment, and retain only segments with at least three tokens. For each retained segment  $s$ , we compute its token overlap with the set of answer tokens:

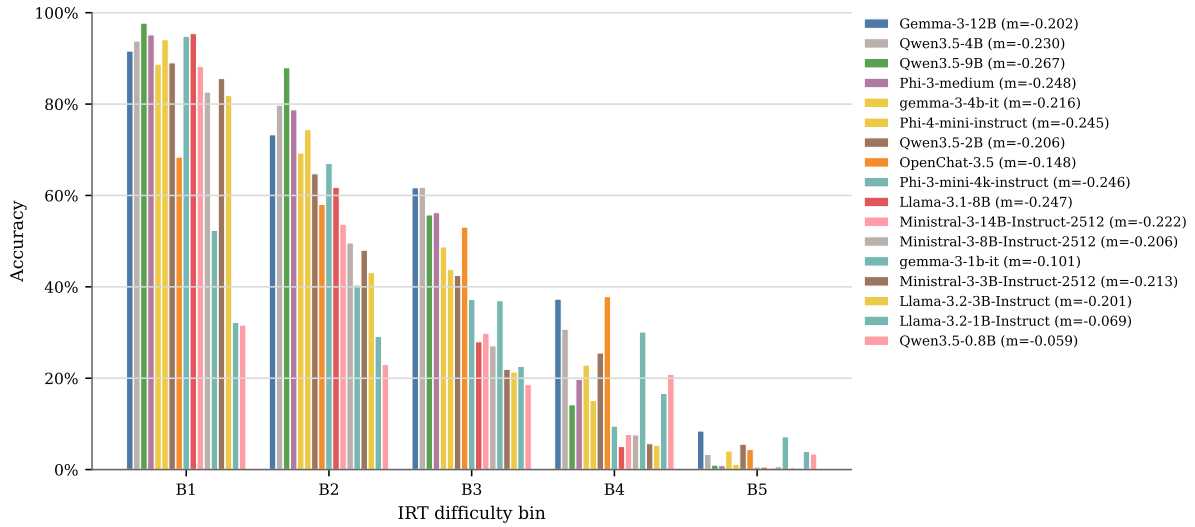
$$\text{Coverage}(s, a_i) = \frac{|U(s) \cap U(a_i)|}{|U(s)|}.$$

A reference segment is considered covered if this value is at least 0.5. The number of missing reference segments is then

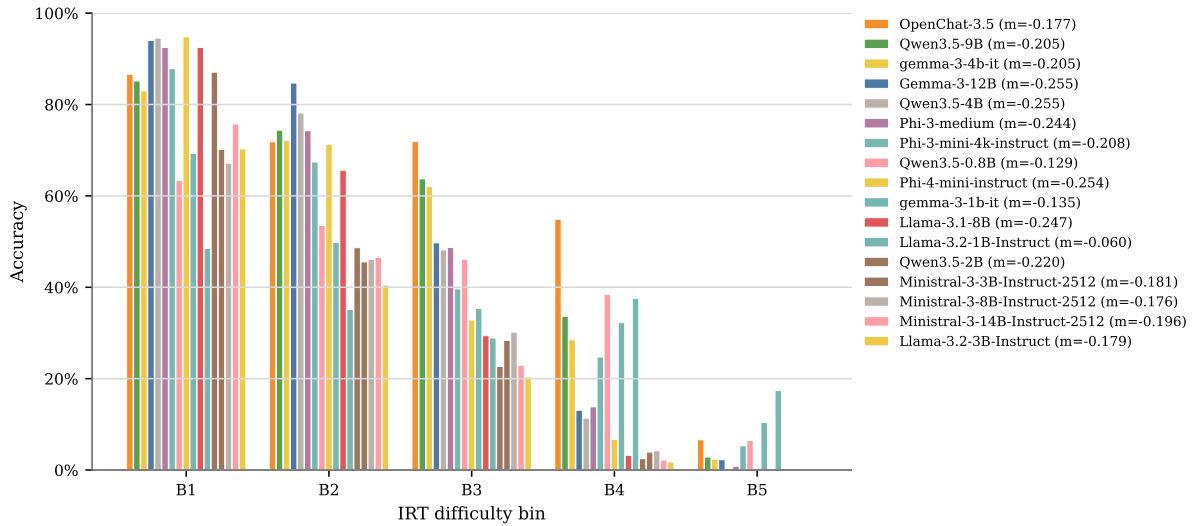
$$\begin{aligned} \text{MissingSegments}(i) &= N_{\text{segments}}(r_i) \\ &\quad - N_{\text{covered}}(r_i). \end{aligned}$$

Higher values indicate that the student answer fails to cover more reference segments.

## A.2 Additional Plots and Tables



(a) SciEntsBank



(b) Beetle

Figure 3: Model accuracy across ordered IRT-based difficulty bins for all evaluated models on (a) SciEntsBank and (b) Beetle. Here,  $m$  denotes the slope obtained by linearly regressing model accuracy on the order of the IRT-based difficulty bins.

LLMs	SciEntsBank			Beetle		
	$\theta$	Acc	Macro-F1	$\theta$	Acc	Macro-F1
google/gemma-3-12b-it	0.805	0.546	0.474	0.487	0.494	0.463
Qwen/Qwen3.5-4B	0.758	0.538	0.434	0.404	0.470	0.314
Qwen/Qwen3.5-9B	0.575	0.508	0.402	0.653	0.518	0.395
microsoft/Phi-3-medium-4k-instruct	0.545	0.499	0.428	0.330	0.464	0.375
google/gemma-3-4b-it	0.367	0.467	0.377	0.515	0.496	0.434
microsoft/Phi-4-mini-instruct	0.309	0.455	0.322	0.084	0.416	0.253
Qwen/Qwen3.5-2B	0.285	0.455	0.350	-0.463	0.322	0.172
openchat/openchat-3.5-0106	0.264	0.446	0.356	0.963	0.577	0.494
microsoft/Phi-3-mini-4k-instruct	0.117	0.416	0.335	0.197	0.450	0.414
meta-llama/Llama-3.1-8B-Instruct	-0.100	0.379	0.224	-0.125	0.386	0.338
mistralai/Ministral-3-14B-Instruct-2512	-0.211	0.359	0.269	-0.645	0.296	0.246
google/gemma-3-1b-it	-0.301	0.337	0.250	-0.075	0.391	0.273
mistralai/Ministral-3-8B-Instruct-2512	-0.349	0.335	0.242	-0.649	0.297	0.282
mistralai/Ministral-3-3B-Instruct-2512	-0.398	0.323	0.181	-0.645	0.298	0.264
meta-llama/Llama-3.2-3B-Instruct	-0.491	0.304	0.148	-0.798	0.266	0.161
meta-llama/Llama-3.2-1B-Instruct	-1.053	0.209	0.144	-0.342	0.331	0.188
Qwen/Qwen3.5-0.8B	-1.122	0.197	0.132	0.109	0.412	0.193

Table 4: Latent grading ability estimates ( $\theta$ ), accuracy, and macro-F1 for LLM judges on SciEntsBank and Beetle under the IRT model.

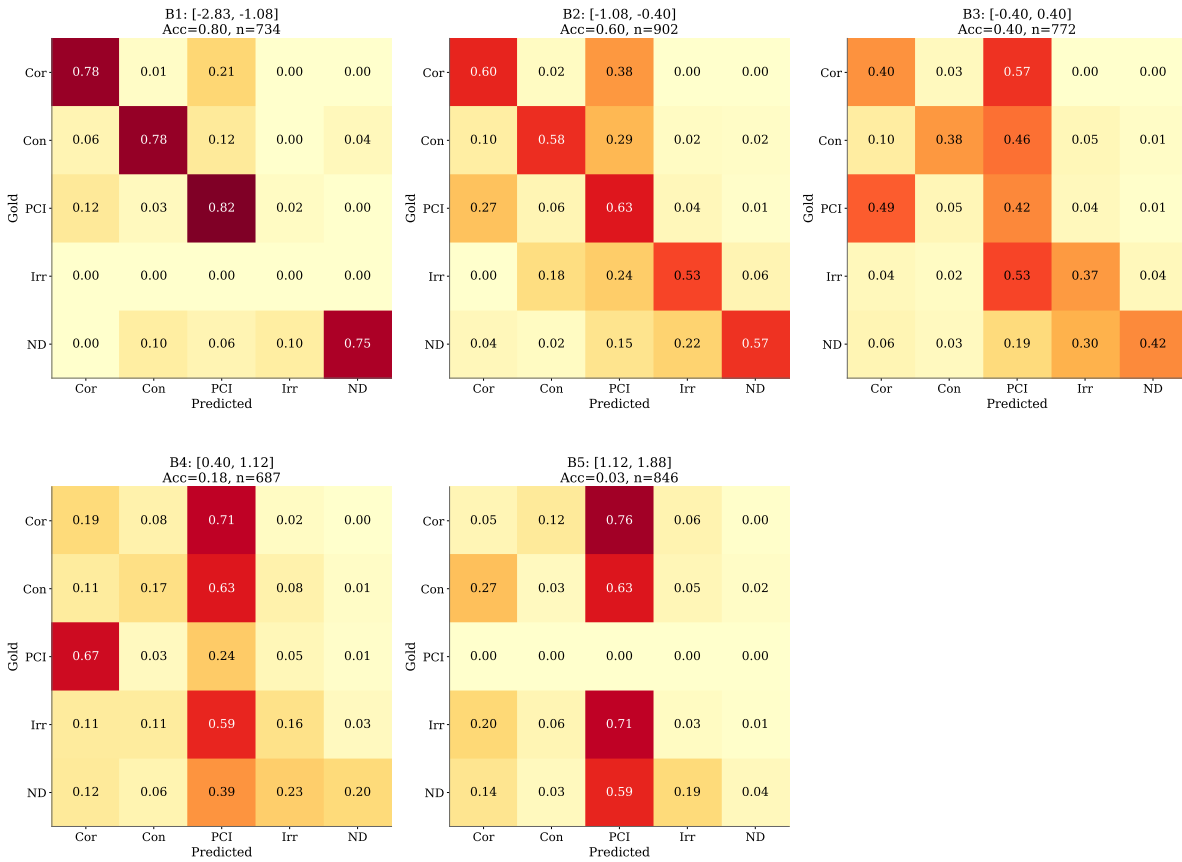


Figure 4: Confusion matrices on Beetle across five bins of IRT-derived response difficulty  $b$ .  $n$  denotes the number of responses in each bin, and Acc denotes overall grading accuracy.