

Math Item Difficulty Prediction with Multimodal Input

Kai North¹, Christopher Ormerod², Alexander Kwako¹, Suhwa Han¹

¹Cambium Assessment ²College Board

Correspondence: Kai.North@cambiumassessment.com

Abstract

Item difficulty prediction relies on the efficient utilization of high-leverage item characteristics. Many items in the domain of mathematics include figures, charts, or other visual stimuli that are challenging to incorporate into difficulty prediction models. Multimodal large language models (LLMs) offer a way to process these visual stimuli in combination with text input, potentially enhancing the success of item difficulty prediction. In this study, we employ several open-source multimodal LLMs to predict the difficulty of math items with visual stimuli from a publicly available data set. We find that multimodal LLMs are capable of predicting math item difficulty.

1 Introduction

Educational assessments often require items of varying difficulty levels in order to adequately measure and differentiate students' proficiency. In large-scale assessment, these difficulty levels are often determined by means of pretests, or field tests. In embedded field tests, new items may be provided to students alongside previously calibrated items. Responses to these new items are then used to determine their overall difficulty for future exams (Yaneva et al., 2024). Pretesting is a necessary yet time-consuming and costly process (Settles et al., 2020). Item difficulty prediction offers a way to limit pretesting requirements by providing initial estimates of key item parameters.

NLP methods and machine learning models can be used to automatically predict item difficulty. For instance, various models ranging from feature-engineered approaches, transformers to more recent LLMs are capable of accurately predicting item difficulty for various text-based items, ranging from multiple-choice to essay-style questions (Yaneva et al., 2024). However, it becomes more challenging to predict item difficulty for items that lack textual context and rely primarily on visual

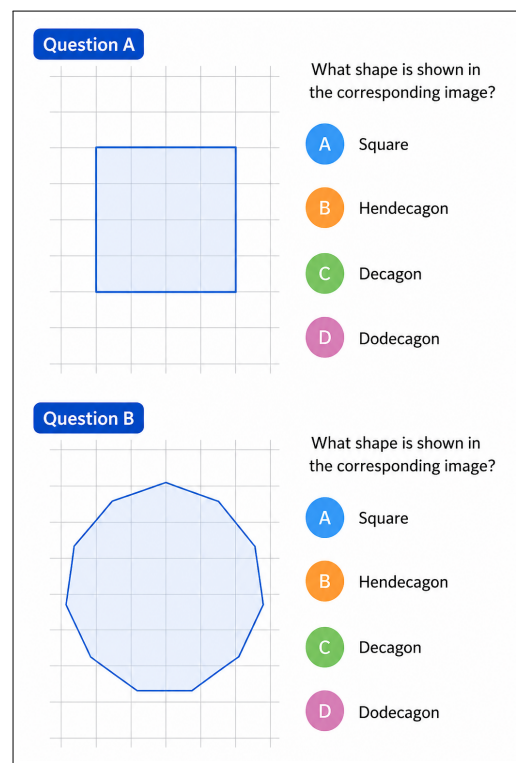


Figure 1: Example math item adapted from NeurIPS 2020 Education Challenge (Wang et al., 2021).

cues or stimuli (Zhu et al., 2025; Acquaye et al., 2026).

Math items may consist of figures, charts, or diagrams that contain little text and cannot be interpreted without their corresponding visual stimuli and possible answers (Wang et al., 2021). The difficulty of the questions shown in Figure 1 cannot be determined accurately using only their stem (question). For example, both questions have the same stem and possible answers, yet their visual stimuli and subsequent perceived difficulty may be different. Question A presents a square and therefore is likely to be considered an easier question for any student in grade 3-10 (aged 8 to 16 years old). This is since a square is a shape learned early within a child's development. Question B, however,

shows a hendecagon, a far less familiar shape for students in this grade range. Furthermore, the additional possible answers (distractors): *Decagon* and *Dodecagon*, are similar shapes to the correct answer: *Hendecagon*, adding to the questions overall difficulty. For this reason, Question B is likely considered more difficult than compared to Question A despite having the same stem and possible answers. As such, their difficulty cannot be accurately predicted using only their text-based features. Instead, their visual stimuli alongside their possible answers are central to their overall difficulty.

This paper explores the use of multimodal LLMs for predicting math item difficulty for grades 3-10 using both an item's text and visual stimuli. Little research has been conducted on the use of these models for predicting the difficulty of math items (Zhu et al., 2025; Acquaye et al., 2026). This paper contributes to the following research question:

RQ. To what extent does the inclusion of visual stimuli in the form of a figure, chart, or diagram impact the performance of automatic math item difficulty prediction.

2 Related Work

Models trained for item difficulty prediction typically infer item difficulty by examining an item's lexical, syntactic, and readability characteristics. These features include lexical richness, word and sentence length, familiarity, sentence structure and cohesion, along with a variety of other psychometric features, for example, grade-level, subject domain, and recorded response time (the time it takes a student to answer a particular item) (Beinborn et al., 2014, 2015; Loukina et al., 2016; El Masri et al., 2017; Zhou and Tao, 2020).

Mojoyinola et al. (2025) trained an XGBoost model on two sets of features obtained either from conventional approaches and LLM-extracted features from GPT-4.1 for math item difficulty prediction. They measured the impact each feature had on their model's performance. Features derived from the item's text were found to be the most influential. These included readability measures: Coleman-Liau and Flesch scores as well as character and syllable count.

Acquaye et al. (2026) investigated the performance of several LLMs, including the Gemma-2, Gemma-3, Llama-3.1, Llama-3.2, and Qwen-2.5 and Qwen-3 series models (of various sizes),

for predicting the item difficulty of 631 multiple-choice math items. They found that directly prompting these LLMs with textual data resulted in very poor performance, with the best Pearson and Spearman correlations reaching only 0.10 and 0.137, respectively, achieved by Qwen-2.5 (32B). However, performance improved substantially when the models were used to simulate student responses across varying ability levels. These simulated responses were then used to estimate item difficulty. This approach yielded a much stronger Pearson correlation of 0.82, obtained from an ensemble of the top-performing LLMs.

Razavi and Powers (2026) also attempted to prompt an LLM directly to predict math item difficulty. They provided GPT-4o with a total of 2564 math items spanning grades Kindergarten (K) to grade 5 assigned with Rasch Item Response Theory (IRT) parameter b difficulty ratings ranging from -5.52 to 3.68 (See Section 3.2 for more information regarding IRT). Their inputted prompt instructed GPT-4o to act as an expert in K-12 math and provide a difficulty estimation based on the item's content. Their best reported RMSE and MAE were 0.74 and 0.60 respectively.

Performance is varied for math item difficulty prediction using text-based LLMs. The aforementioned studies suggest that prompting an LLM results in low Pearson and Spearman correlations and RMSE and MAE scores which could be improved when compared to other item difficulty prediction tasks (Yaneva et al., 2024). These studies also report that textual features are highly influential (Mojoyinola et al., 2025; Razavi and Powers, 2026) and higher performance can be achieved through indirect methods: student simulation (Acquaye et al., 2026). However, these studies do not investigate the impact visual stimuli: figures, charts, and diagrams, have on overall performance.

2.1 Visual Stimuli for Predicting Difficulty

Little research pertains to the use of multimodal LLMs for math item difficulty prediction. Zhu et al. (2025) were interested in distinguishing between easy and hard math items without model fine-tuning. They conducted some preliminary experiments into the use of multimodal LLMs: Qwen2.5-VL and InternVL3. They tested these models at binary classification. Models were presented with math items and tasked with labeling them as either easy (0) or hard (1). They report noticing both multimodal LLMs showing a clear distinction between

the two class types. They leveraged the hidden representations of these LLMs within a Markov chain to predict final difficulty estimations achieving Macro-F1s of 80.68 and 75.98 for Qwen2.5-VL and InternVL3 respectively.

Within this study, we extend prior work on the use of multimodal LLMs for math item difficulty prediction. However, unlike [Zhu et al. \(2025\)](#), we approach math item difficulty prediction as a regression-based task, training several multimodal LLMs to predict a continuous item difficulty parameter (b) obtained through IRT calibration (Section 3.2).

3 Methods

3.1 Dataset

A total of 947 items were included in a publicly available dataset provided by Task 3 of the NeurIPS 2020 Education Challenge ([Wang et al., 2021](#)).¹ Task 3 asked participating teams to design models to predict “the quality of a question, as defined by a panel of domain experts (experienced teachers)” ([Wang et al., 2021](#)). Each math item contains a figure, graph or other illustration directly in reference to the provided question. Questions were multiple-choice questions (MCQs) spanning subjects such as “Algebra”, “Data and Statistics”, and “Geometry and Measure” ([Wang et al., 2021](#)) and were designed for students from grades 3-10. The dataset also provides student responses for each question; not all students answered all questions. To obtain difficulty parameter estimates (our “gold labels”), we fit a Rasch model similar to [Razavi and Powers \(2026\)](#). It is important to note, that there is a lack of available math item datasets that contain both visual stimuli and human annotated difficulty ratings, hence why we chose to conduct our own IRT calibration.

3.2 IRT Calibration - Difficulty Parameter (b)

IRT encompasses a family of models that estimate some latent trait(s), e.g. students’ math proficiency, on the same scale as item characteristics, e.g. item difficulty ([Razavi and Powers, 2026](#)). We refer to students’ math proficiency as θ and item difficulty as b , common abbreviations in the field. When θ scale is centered at 0, a θ of 0 indicates average proficiency within the sample, a negative θ indi-

cates lower than average proficiency, and positive higher than average. Under a simple dichotomous model such as a Rasch model, a b of 0 indicates that a student of proficiency $\theta = 0$ will answer the question correctly 50% of the time.

Given the small and sparse nature of the dataset, we select a simple IRT model, the Rasch model. We estimated difficulty parameters using marginal maximum likelihood estimation implemented in the `mirt` package (Version 1.45) in R ([Chalmers, 2012](#)), assuming a normally distributed latent variable θ with its mean fixed at 0 for identification. The Rasch model yielded stable estimates of standard error of difficulty parameters. To improve difficulty estimates, we removed items with fewer than 50 student responses (resulting in 126 items being removed). We obtained b parameter estimates for 821 math items. We divided these items into train and test splits of 656 (80%) and 165 (20%) respectively.

Text Only - Prompt

Question: <text>What shape is shown in the corresponding image? A. Square, B. Hendecagon, C. Decagon, D. Dodecagon<text>

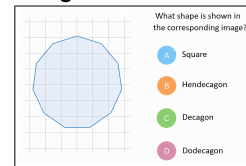
Instruction: Estimate the item difficulty (IRT B parameter) for the provided question. Output one real number with up to two decimal places.

0 represents average difficulty.
Negative values indicate easier items.
Positive values indicate more difficult items.

Do not include words, punctuation, or explanations.

Image & Text - Prompt

<image>



<image>

Question: <text>What shape is shown in the corresponding image? A. Square, B. Hendecagon, C. Decagon, D. Dodecagon<text>

Instruction: Estimate the item difficulty (IRT B parameter) for the provided question. Output one real number with up to two decimal places.

0 represents average difficulty.
Negative values indicate easier items.
Positive values indicate more difficult items.

Do not include words, punctuation, or explanations.

Table 1: Prompt types used for training and evaluation.

¹Data is currently hosted by Eedi at <https://www.eedischool.com/projects/neurips-education-challenge>

3.3 Models

We trained and evaluated several multimodal LLMs on the above math items for regression-based prediction of their IRT difficulty parameter (b), hereby referred to as simply IRT b . These models include Qwen3.5-9B (Alibaba Cloud Qwen Team, 2025b), Qwen2.5-VL-7B (Alibaba Cloud Qwen Team, 2025a), Qwen2-VL-7B (Alibaba Cloud Qwen Team, 2024) and Llama-3.2-11B-Vision (Meta AI, 2025). Each model was fine-tuned on the following prompts (Table 1).

The first prompt provides the text associated with the math item, including only its stem and possible answers, and asks the model to estimate its difficulty by outputting a single numerical IRT b . In addition to the stem and possible answers, the second prompt also provides the math item’s corresponding visual stimuli in the form of a single figure, chart, or diagram, and likewise tasks the model to predict the math item’s IRT b .

Training was conducted over 10 epochs with a learning rate of $2e-4$ on a NVIDIA L40S GPU. Each multimodal LLM was fine-tuned using a custom regression-head. This entailed augmenting the model loss function with a feed-forward regression module applied to the final hidden state representation of the last token, producing a scalar prediction of IRT b . The overall objective combined the standard language modeling loss with a mean squared error (MSE) loss on the regression targets, jointly optimized via a weighted sum. This is represented as:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{MSE}} \quad (1)$$

where:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (\hat{b}_i - b_i)^2 \quad (2)$$

and $\lambda = 0.1$ is a weighting coefficient that controls the contribution of the regression objective relative to the language modeling loss. Here, $\mathcal{L}_{\text{LM}}$ denotes the standard token-level cross-entropy loss computed over the predicted output sequence, while $\mathcal{L}_{\text{MSE}}$ measures the discrepancy between the predicted scalar regression value for IRT $\hat{b}_i$ and the gold IRT b_i for each sample in the batch of size N . In this way, our multimodal LLMs were adapted for regression and capable of making IRT b predictions.

Multimodal LLM performances were compared against two additional text-only models: ModernBERT (Warner et al., 2025) and Electra (Clark et al., 2020). These models were also trained for regression and the prediction of each math item’s IRT b . They were trained and evaluated on the same sets of items. Training was over 10 epochs with a learning rate of $2e-4$. These additional models were chosen to provide further text-only baselines for assessing the benefits of including visual stimuli for math item difficulty prediction.

3.4 Metrics

Evaluation was conducted using mean absolute error (MSE), root mean squared error (RMSE) and Pearson correlation. These evaluation metrics are standard for regression-based tasks and have been frequently used for item difficulty prediction (Yaneva et al., 2024; Acquaye et al., 2026; Razavi and Powers, 2026). They are defined as follows:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2 \quad (3)$$

where $\hat{y}_i$ and y_i denote the predicted and gold values for the i -th sample, respectively, and N is the total number of samples. MSE measures the average squared difference between predictions and gold values, placing higher penalty on large errors.

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (4)$$

where RMSE is the square root of MSE and provides an error metric in the same units as the target variable, making it more interpretable.

$$r = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}} \quad (5)$$

where r denotes the Pearson correlation coefficient, $\bar{\hat{y}}$ and $\bar{y}$ are the mean predicted and gold values, respectively. Pearson correlation measures the linear relationship between predictions and gold values, with values closer to 1 indicating stronger positive correlation.

4 Results

The majority of multimodal LLMs were found to perform better when being trained on a combination of item text and visual stimuli than compared

Model	Text Only			Image & Text		
	MSE	RMSE	Pearson	MSE	RMSE	Pearson
Qwen2.5-VL-7B	0.35	0.59	0.68	0.27	0.52	0.77
Qwen3.5-9B	0.30	0.55	0.74	0.28	0.53	0.76
Qwen2-VL-7B	0.35	0.59	0.68	0.32	0.56	0.73
Llama-3.2-11B-Vision	0.36	0.60	0.68	0.37	0.61	0.67
ModernBERT-base	0.44	0.66	0.58	—	—	—
Electra-base	0.45	0.67	0.58	—	—	—

Table 2: Regression performance for predicting item difficulty (IRT b parameter) under Text Only and Image & Text modalities.

to text only for math item difficulty prediction. Table 2 shows that Qwen2.5-VL was the best performing model (in bold) having achieved an MSE of 0.27, RMSE of 0.52 and a Pearson correlation of 0.77 when trained on the image and text prompt. Qwen2.5-VL achieved a slightly worst MSE of 0.35, RMSE of 0.59 and a Pearson correlation of 0.68 when trained on the text only prompt. This finding was observed for all Qwen variants with an average improvement of -0.04 MSE, -0.04 RMSE, and +0.05 Pearson correlation found across models when trained on both text and visual stimuli compared to when being trained on the text-only prompt.

Interestingly, Llama 3.2-11B-Vision did not show any benefit from the inclusion of visual stimuli. Instead, little-to-no change in performance was observed when Llama 3.2-11B-Vision was trained on either image and text or text only prompts. The greater size of this model, being 11B parameters, may explain why it does not show improvement from the inclusion of images. Llama 3.2-11B-Vision may already be capable of extracting sufficient information about item difficulty from the item’s text alone, leaving limited opportunity for additional visual information to provide any meaningful insight. Alternatively, Llama 3.2-11B-Vision may be less effective at extracting visual features than compared to Qwen-based models from math items.

ModernBERT and Electra provide additional text-only baselines. These models were observed to perform worse than all multimodal LLMs when fed either prompt type (Table 2). ModernBERT and Electra achieved nearly identical performance, attaining MSE values of 0.44 and 0.45, RMSE values of 0.66 and 0.67, and Pearson correlations of 0.58. These results suggest that traditional transformer encoders are less effective at capturing item difficulty when presented with math

items containing limited text. On the other hand, our multimodal LLMs are able to make use of their large-scale pre-training to infer difficulty from less text and utilize visual stimuli unavailable to text-only approaches.

5 Discussion and Conclusion

In this study, we demonstrate that images containing figures, charts, or diagrams can be leveraged via multimodal LLMs to potentially improve math item difficulty prediction for some models. Qwen variant models appeared to benefit from the inclusion of visual stimuli whereas Llama 3.2-11B-Vision showed no improvement. Multimodal LLMs can leverage this additional visual information, whereas text-only models must rely solely on textual input; hence, multimodal LLMs are able to make predictions based on a more comprehensive set of item characteristics.

Previous research has shown mixed performances when using text-only LLMs for math item difficulty prediction. Acquaye et al. (2026) reported Pearson and Spearman correlations of 0.1 and 0.137 respectively for Qwen-2.5. Razavi and Powers (2026) obtained an RMSE of 0.74 and an MAE value of 0.60 from GPT-4o. These studies and others have attempted to improve upon their performances by using LLMs to either mimic student responses or to provide features which can later be used to predict math item difficulty by another model. Unlike prior research, this paper shows that math item difficulty prediction can be achieved directly through the use of multimodal LLMs.

For future research, we aim to refine our prompt with the aim of improving performance further. We suspect that the inclusion of each item’s rubric or rationale within the prompt may enhance overall performance. Additionally, experimentation with larger LLMs is required given the performance

of Llama-3.2-11B-Vision. Lastly, we also aim to provide further analysis investigating model performance per grade level. We hypothesize that grade level may impact the usefulness of visual stimuli for predicting math item difficulty.

6 Limitations

This study is limited by the significant time and computational resources required to train and evaluate multimodal LLMs. Due to these constraints, we were only able to include a subset of available models, and given more time and resources, we aim to incorporate additional models in future work to provide a broader comparison of overall performance. Another limitation arises from the availability of datasets containing both math images and reliable difficulty labels. Such datasets are scarce, which restricts the diversity of item types and complexity levels that our models can be evaluated on. Furthermore, the limited size of these datasets may impact the generalizability of our findings.

Another possibility for incorporating visual stimuli into text-only LLMs would be to use multimodal LLMs to describe the visual elements within a given math item image. This description could then be fed into a text-only based LLM. Although this could result in comparable performance, this approach would also add an additional step within the item difficulty pipeline. Moreover, this approach may not capture key aspects of the stimuli; by contrast, multimodal LLMs predict item difficulty directly from math images.

References

- Christabel Acquaye, Yi Ting Huang, Marine Carpuat, and Rachel Rudinger. 2026. Take out your calculators: Estimating the real difficulty of question items with llm student simulations. *ArXiv*, abs/2601.09953.
- Alibaba Cloud Qwen Team. 2024. Qwen2-vl-7b. <https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct>.
- Alibaba Cloud Qwen Team. 2025a. Qwen2.5-vl-7b: Multimodal large language model. <https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct>.
- Alibaba Cloud Qwen Team. 2025b. Qwen3.5-9b. <https://huggingface.co/Qwen/Qwen3.5-9B>.
- Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2014. Predicting the difficulty of language proficiency tests. *Transactions of the Association for Computational Linguistics*, 2:517–530.
- Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2015. Candidate evaluation strategies for improved difficulty prediction of language tests. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–11.
- R. Philip Chalmers. 2012. *mirt: A multidimensional item response theory package for the r environment*. *Journal of Statistical Software*, 48(6):1–29.
- Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of ICLR*.
- Yasmine H El Masri, Steve Ferrara, Peter W Foltz, and Jo-Anne Baird. 2017. Predicting item difficulty of science national curriculum tests: The case of key stage 2 assessments. *The Curriculum Journal*, 28(1):59–82.
- Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. 2016. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3245–3253.
- Meta AI. 2025. Llama 3.2 11b vision. <https://huggingface.co/meta-llama/Llama-3.2-11B-Vision>.
- Mubarak Mojoyinola, Olasunkanmi James Kehinde, and Judy Tang. 2025. Enhancing item difficulty prediction in large-scale assessment with large language model. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Works in Progress*, Wyndham Grand Pittsburgh, Downtown, Pittsburgh, Pennsylvania, United States. National Council on Measurement in Education (NCME).
- Pooya Razavi and Sonya Powers. 2026. *Estimating item difficulty using large language models and tree-based machine learning algorithms*. *Preprint*, arXiv:2504.08804.
- Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8:247–263.
- Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Jordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E. Turner, Richard G. Baraniuk, Eedi Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. 2021. Results and insights from diagnostic questions: The neurips 2020 education challenge. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 191–205. PMLR.
- 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)*, Vienna, Austria. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Ya Zhou and Can Tao. 2020. Multi-task BERT for problem difficulty prediction. In *2020 International Conference on Communications, Information System and Computer Engineering (CISCE)*, pages 213–216.

Yubo Zhu, Dongrui Liu, Zecheng Lin, Wei Tong, Sheng Zhong, and Jing Shao. 2025. The LLM already knows: Estimating LLM-perceived question difficulty via hidden representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, Suzhou, China. Association for Computational Linguistics.