

# Calibrated, Interpretable Automated Scoring with LLM Likelihoods

**S. Thomas Christie**

Renaissance Philanthropy  
thomas.christie@renphil.org

**Markus Hauru**

Renaissance Philanthropy  
markus.hauru@renphil.org

**Anna N. Rafferty**

Carleton College  
arafferty@carleton.edu

## Abstract

Some learning is difficult to capture using selected-response items and is better revealed through tasks involving writing and other rich behavior. Automated scoring of such work products has proven difficult, and typical scoring methods do not support principled degree-of-belief quantification for individual scores or evidence accumulation across tasks. We investigate whether generative models can be used to quantify the evidence provided by student writing. In particular, we use conditional token likelihoods produced by LLMs to score student answers in the SciEntsBank dataset. We find that, for predicting expert labels, zero-shot likelihood-based methods provide scores with accuracy comparable to, and calibration much better than, LLM ‘judges’. Well-calibrated scores support response-level and student-level uncertainty quantification, enabling more trustworthy, useful, and interpretable automated scoring.

## 1 Introduction

In educational assessment, scoring processes often reduce rich cognitive and behavioral outputs to a single value: a half-page of mathematical notation might be coded as ‘incorrect’, or an explanation of a scientific phenomenon reduced to a 4/5. Scores summarize the adherence of a final work product to a reference or rubric but rarely record the scorer’s uncertainty about whether the assigned score is correct. In effect, scores are treated as outputs of an infallible classifier.

A score should reflect a claim about a student’s performance on a particular task (Mislevy et al., 2003; DeCarlo et al., 2011), and should express the *degree of evidence* that the observed behavior supplies for the claim. To achieve this, we need a scoring mechanism that produces a probability distribution over claims, rather than categorical labels. Well-calibrated probabilities support evidence accumulation, determinations of required assessment

lengths, and decisions about which responses warrant human review. Because probabilities reflect small accumulations of evidence that categorical labels mask, computing them incrementally can even indicate which components of a student’s response provide evidence for a claim, a critical signal for interpretability and bias analysis.

In this paper, we evaluate and compare the performance of three scoring methods powered by large language models (LLMs) on responses in the SciEntsBank dataset. We evaluate scoring performance along three axes: accuracy at reproducing human labels, prediction calibration, and prediction interpretability. While classification accuracy is important, the main focus of our analysis is confidence calibration. A well-calibrated scorer is one whose confidence in its score is an accurate measure of the probability that the score is correct. While typical LLM-powered scoring methods use LLMs to generate text labels for scores, we show that two methods that leverage LLM token likelihoods achieve much better calibration than a judge approach while maintaining similar accuracy.

## 2 Background

In recent work, we argued that because LLMs and other generative methods compute conditional likelihoods natively, they may be able to operate as measurement equations for student-produced text (Christie et al., 2026). If so, student skills and abilities may be directly inferable from rich, naturalistic behavior using probabilistic machinery. In the proposed approach, a student’s attribute is represented in natural language as a prompt component  $Y$ , with various attribute values (say, ‘proficient’ versus ‘not proficient’) encoded as  $Y_1, \dots, Y_n$ . An LLM  $M$  is used to compute, for each attribute value, the likelihood of the student output  $X$  given the item description  $D$ :

$$p_M(X \mid Y_i, D) \text{ for all } i = 1, \dots, n$$

From these values, a posterior distribution over the mutually exclusive attribute values  $Y_1, \dots, Y_n$  can be computed. This method draws on an approach called the *noisy channel model* (henceforth *channel*; Min et al., 2022), and has been explored for text classification but had not previously been applied to educational measurement. The method is named after a formalism in information theory in which a message is transmitted over a noisy channel and observed by the receiver, who is tasked with inferring the most likely source message.

In a similar approach, the LLM likelihood model can be used to compute the conditional probabilities of a label directly, namely,

$$p_M(Y_i | X, D) \text{ for all } i = 1, \dots, n$$

which are then normalized. Following Min et al. (2022), we refer to this as the ‘direct’ method. Like the channel method, the direct method derives a posterior over labels from LLM token probabilities.

In the common ‘LLM-as-a-judge’ paradigm (henceforth ‘judge’; Gu et al., 2026), a model is prompted with a rubric and a student-written text and asked to generate a score or label. While this approach can produce strong classification results (e.g., Naismith et al., 2023), scores are not accompanied by principled estimates of score uncertainty. Several methods for deriving confidence estimates from LLM judgments have been proposed, including simply asking the LLM to state its confidence alongside its score (Tian et al., 2023). These uncertainty estimates are themselves generated text rather than readouts of the model’s conditional likelihoods; they reflect a model’s self-reported belief, which may not be consistent with its underlying probabilities.

Automated scoring methods can be broadly categorized as supervised (informed by many examples) or zero-shot. Supervised scoring methods can achieve high accuracy by training specialized models (e.g., Kaya and Cicekli, 2024) or fine-tuning LLMs (Karim et al., 2025). While the literature reporting calibration is limited, some supervised approaches also achieve strong calibration (Kapoor et al., 2024; Lade et al., 2026). Despite this, we focus on zero-shot scoring methods for two reasons. First, supervised methods require substantial expert-labeled training data, which is often costly or unavailable. Second, we envision a future in which students can flexibly choose how to demonstrate their learning, and downstream scor-

ing and measurement are adaptable enough to produce valid and reliable inferences regardless. We view the current work as a step toward that future.

### 3 Data

We evaluate channel, direct, and judge methods on the SciEntsBank dataset (Dzikovska et al., 2013), which consists of 3rd–6th grade student responses to short-answer science questions. We limit analysis to the “unseen answers” (test\_ua) portion of the dataset, which has 540 responses. The human-annotated ground truth differentiates among types of incorrect answers. Following the 2-way classification task (Dzikovska et al., 2013), we map the five labels to correct-versus-incorrect, turning the problem into a binary classification task.

### 4 Methods

Consider an item description  $D$ , a student’s answer to it  $X$ , and labels  $Y_C$  and  $Y_I$  indicating whether the student’s answer is judged correct or incorrect, respectively.  $D$  may include other context, such as a reference answer or a scoring rubric, depending on the data. In our experiments,  $D$  is composed of the item question, the reference answer, and brief descriptions of the two labels. We wish to compute the posterior probabilities of the labels given the student’s answer:  $p(Y_i | X, D)$ . We use three different methods to elicit these score probabilities from an LLM and compare the resulting predictions to human labels. Our experiments focus on small and mid-sized open-weight LLMs as base models; the judge can alternatively use a closed-weight model, so we include a comparison to OpenAI’s GPT models for reference.

#### 4.1 Score elicitation approaches

Of our three score elicitation approaches, the two likelihood-based methods require an LLM to produce the probability of each possible next token (e.g., a word or word piece), given the preceding text, rather than generate new text. We denote these next-token probabilities by  $p_M$  for a model  $M$ .

The probability assigned by  $M$  to a longer text made of tokens  $t_0$  to  $t_N$ , denoted  $p_M(t_{\leq N})$ , is the product of the probabilities assigned to each consecutive token, conditioned on all preceding tokens:

$$p_M(t_{\leq N}) = p_M(t_0) \prod_{j=1}^N p_M(t_j | t_{\leq j-1})$$

| Property                           | Judge | Direct | Channel |
|------------------------------------|-------|--------|---------|
| Native per-token evidence          | ×     | ×      | ✓       |
| Immune to surface-form competition | ×     | ×      | ✓       |
| Runs over closed (API-only) models | ✓     | ×      | ×       |
| Supports “thinking” models         | ✓     | ×      | ×       |

Table 1: Qualitative comparison of the three uncertainty-aware LLM-based scoring methods we study. ✓/× indicate whether a method provides the property.

Access to next-token probabilities typically requires an open-weight LLM, since most closed-model APIs do not expose next-token probabilities for all tokens.

**Direct method:** In the direct method, we prepare a text that includes the item description  $D$  and student answer  $X$ . We use the LLM to compute  $p_M(Y_i | X, D)$  for  $i \in \{C, I\}$  and normalize the values, treating the result as the posterior distribution over labels. Some scaffolding is required to make the text read like natural language for the LLM; an abbreviated example prompt is:

A student was asked "How can one reduce the current in a circuit?" The student’s answer was "by increasing the resistance". The student’s answer is

The LLM  $M$  is queried for the probabilities that the next token in this text is ‘correct’ or ‘incorrect’.

To mitigate the bias whereby an LLM may favor one label or the other regardless of the student’s answer, we apply the content-free calibration from Zhao et al. (2021). In particular, we apply this calibration question-by-question, by using a prompt that includes the question and the reference answer, and that replaces the student answer with the string [REDACTED]; see Zhao et al. (2021) for details.

**Noisy channel model:** In the noisy channel model, the order of the student answer and label is inverted. The prepared prompt includes  $D$  and  $Y_i$ , and we compute the probability  $p_M(X | Y_i, D)$ . As noted above, this is structurally similar to a psychometric measurement equation. We compute the posterior probability using Bayes’ theorem:

$$p(Y_i | X, D) = \frac{p_M(X | Y_i, D) p(Y_i | D)}{\sum_{j \in \{I, C\}} p_M(X | Y_j, D) p(Y_j | D)}.$$

Here  $p(Y_i | D)$  is the prior over answer correctness, which we set to be uniform. The abbreviated example prompt for the channel method is:

A student was asked "How can one reduce the current in a circuit?" The student’s answer was correct. The student’s answer was

and we use  $M$  to compute the conditional likelihood of the continuation ‘by increasing the resistance’. The procedure is repeated by replacing ‘correct’ with ‘incorrect’ in the prompt, and the two likelihoods are normalized.

**LLM-as-a-judge:** The LLM-as-a-judge (judge) method prompts an LLM for both a label and a verbalized confidence rather than using likelihood readouts directly. With the same abbreviated example as above, the prompt is:

A student was asked "How can one reduce the current in a circuit?" The student’s answer was "by increasing the resistance". Is the student’s answer correct or incorrect? Answer with JSON: "label": "<correct or incorrect>", "confidence": "<probability between 0.0 and 1.0 that the label is correct>".

Note that the judge method should be used with an instruction-tuned LLM, whereas the direct and channel methods work best without instruction fine-tuning; in our experiments, the judge uses each model’s instruction-tuned variant while the channel and direct methods use the corresponding base model.

## 4.2 Method affordances

Method affordances are summarized in Table 1. The channel method natively provides per-token evidence, since it accumulates its score probabilities by computing and comparing the conditional likelihoods of each subsequent student-written token. Similar per-token results can be obtained for the direct and judge methods using prompts where only a partial student answer is shown (see Figure 4). However, this requires rerunning the scoring

Scoring methods on 2-way SciEntsBank (test\_ua, n = 540)

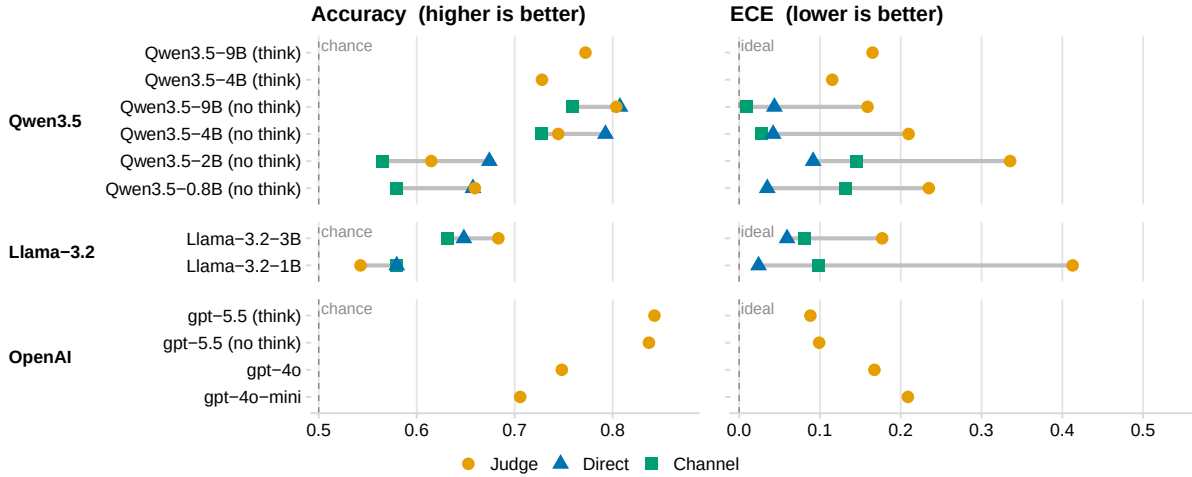


Figure 1: Scoring methods on 2-way SciEntsBank (test\_ua split,  $n = 540$ ), evaluated with various models. On the left accuracy, on the right 10-bin Expected Calibration Error (ECE). The direct method uses the content-free calibration correction (Zhao et al., 2021). Results typically improve with model size for all methods, and accuracies are broadly similar across them when using the same base LLM. Calibration error, however, is far larger for the LLM-as-judge’s self-reported confidence than with the LLM-likelihood-based approaches. The frontier-model judge with GPT-5.5 is superior in accuracy, though still behind the direct method and the channel method with Qwen3.5-9B in calibration error. Our closed-model judge results are broadly in line with those reported by others (Kortemeyer, 2024).

for every token in the answer, whereas the channel method captures this information in one pass.

The channel method is also the only method immune to surface-form competition, a known issue in which probability mass is split across multiple surface forms of the same label, diluting the signal. Methods for mitigating it exist (Holtzman et al., 2021), but are imperfect and complicate implementation.

Both channel and direct methods require open-weight models, so the judge is the only option when only closed-weight models are available.

Finally, LLMs that use thinking traces can, in principle, be used with the direct and channel methods, but are a more natural fit for LLM-as-a-judge. We have not found a way to insert thinking traces into the channel and direct methods without biasing their confidence estimates.

## 5 Results

Score-prediction results for SciEntsBank are shown in Figure 1.<sup>1</sup> All three scoring methods exhibit comparable accuracy for size-matched models, with the channel method performing slightly worse than the other two. The base LLM affects accuracy

<sup>1</sup>Source code for results is at [github.com/RenaissancePhilanthropy/Calibrated-LLM-scoring-AIME-2026](https://github.com/RenaissancePhilanthropy/Calibrated-LLM-scoring-AIME-2026).

more than the scoring method. The small, open models we evaluated are competitive with GPT-4o in accuracy but are outperformed by GPT-5.5.

To measure calibration of confidence estimates, we use Expected Calibration Error (ECE; Naeini et al., 2015; Guo et al., 2017), defined as  $\sum_{k=0}^{K-1} \frac{|B_k|}{N} |\text{acc}(B_k) - \text{conf}(B_k)|$ , where the predictions from the method have been split into  $K$  equal-width confidence bins  $B_k$  over  $[0, 1]$  based on the probability assigned to the top-scoring label, and  $\text{acc}(B_k)$  and  $\text{conf}(B_k)$  are the average accuracy and confidence within that bin. We use  $K = 10$  bins. As shown on the right in Figure 1, in size-matched comparisons, the channel and direct methods exhibit better-calibrated probability estimates than LLM-as-a-judge.

The reliability plots in Figure 2 provide more detail about prediction calibration for each method, using the Qwen3.5-9B model. The channel method uses the full range of possible confidences and demonstrates good calibration: when the model says it is 60% sure of its answer, it is correct approximately 60% of the time. The direct method is similar, if slightly worse. In contrast, the judge uses only a few stereotypical confidence ratings (e.g., 95% and 100%) and is systematically overconfident.

Figure 3 shows coverage and accuracy when

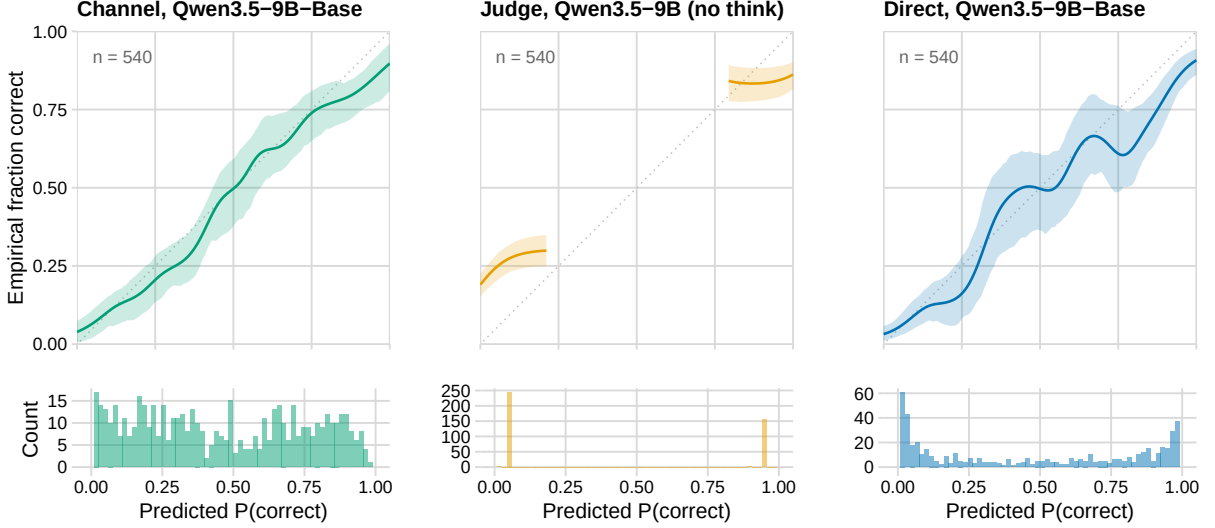


Figure 2: Top: Predicted probability that a student answer is correct versus how often the answer actually was correct. Each sample contributes one point at its predicted  $P(\text{correct})$ ; the curve is a Gaussian-smoothed ( $\sigma = 0.05$ ) empirical frequency of the gold label being *correct*, with a bootstrapped 95% confidence band, and the dotted diagonal marks perfect calibration. Bottom: The same data, showing its distribution along the predicted  $P(\text{correct})$ .

varying a confidence threshold for each of the scoring methods on SciEntsBank, again using Qwen3.5-9B. To produce this plot, for each method we filter to keep only predictions where the model’s confidence is above a chosen threshold. We compute the accuracy within this high-confidence subset and the fraction of all predictions it contains (coverage). We then vary the confidence threshold from 0 to 1, and Figure 3 reports, for each required accuracy level, the largest coverage achievable at that accuracy. In practice, a user may choose to trust high-confidence model predictions and route low-confidence scores to a human scorer. While direct and judge methods have roughly equivalent overall accuracy, the better-calibrated direct method allows responses to be ranked by trustworthiness, supporting downstream decisions about when and whether to rely on model outputs.

We explore score interpretability by investigating whether the resulting score can be attributed to specific tokens in the student answers. While a comprehensive analysis is beyond the scope of this work (and an appropriate human-labeled dataset is unavailable for SciEntsBank), we use example item responses as case studies. In Figure 4, we compute the per-token evidence and resulting posterior probability distribution for two question/response pairs. In the channel method, the log-likelihood ratio after token  $j$  (equal to the posterior log-odds under a uniform prior) is computed directly over the prefix

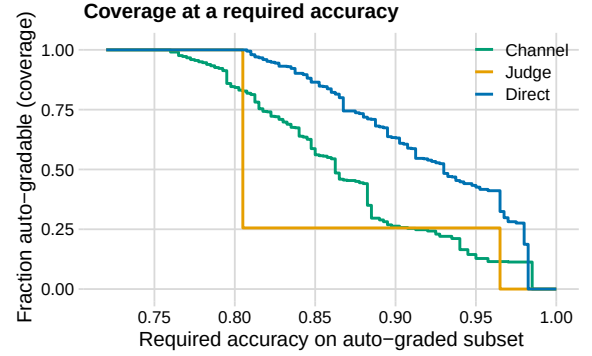


Figure 3: Selective prediction on 2-way SciEntsBank; predictions are from Qwen3.5-9B. For each required accuracy level (horizontal axis), the curve shows the largest fraction of responses that can be auto-scored while meeting that accuracy on the auto-scored subset (vertical axis).

$X_{1..j}$ :

$$\text{LLR} = \log \left( \frac{p_M(X_{1..j} \mid Y_C, D)}{p_M(X_{1..j} \mid Y_I, D)} \right)$$

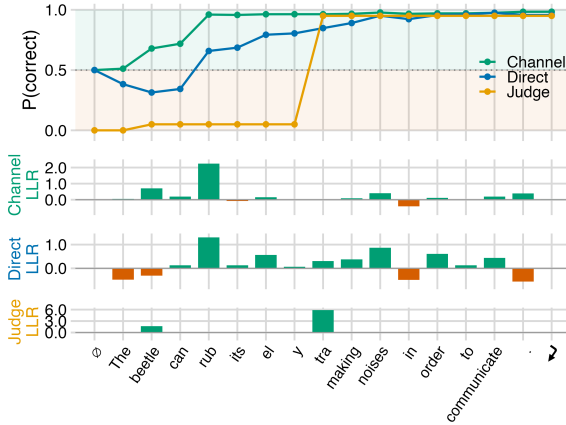
For the direct and judge methods, classifications are made using repeated evaluations of partial (prefix) student responses with the remainder redacted, following the content-free prompt of Zhao et al. (2021). To illustrate, in the first direct-method example, the following prompt would be used to score a partial response:

A student was asked "How can one reduce the current in a circuit?" The

**Token-level interpretability:  
channel (native) vs direct and judge (prefix-reveal)**

Q: One function of the bess beetle's elytra is protection.  
What is another function?

Ref: The elytra are used to make sounds.



Q: What will happen to the motor if the light bulb burns out?  
Ref: The motor will continue to run (still has a pathway to the D-cell).

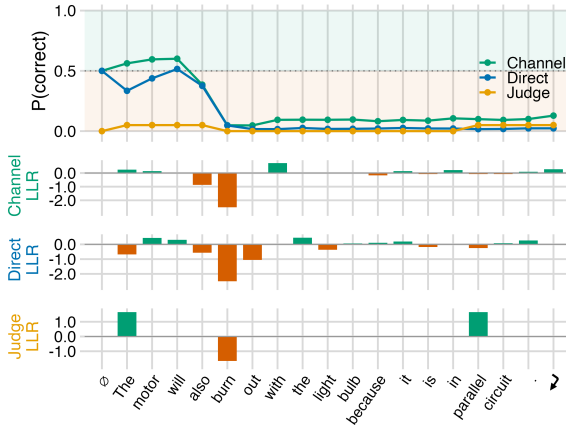


Figure 4: Token-by-token scoring with Qwen3.5-9B on two SciEntsBank question/response pairs. For each example, the top plot overlays each method’s cumulative probability that the student answer is correct as the answer is revealed token by token; the dotted line is the 0.5 decision boundary. The three bar plots show each token’s log-likelihood-ratio contribution (the change in cumulative LLR), i.e., whether it pushed that method’s verdict toward correct (up) or incorrect (down).

student’s answer was "by increasing [REDACTED]". The student’s answer is

Both the channel and direct methods show gradual accumulation of evidence throughout the course of the student text, and words with strong probability changes are intuitively indicative of student answer correctness. In contrast, the judge method collects evidence in bursts, a phenomenon that reflects the coarse, clustered confidence values revealed in Figure 2.

## 6 Discussion

When used in educational assessment, automated scoring is one step in a process of producing valid and reliable estimates of student learning. While most scoring encodes adherence to a rubric, we argue that scores should encode the *degree of evidence* regarding a claim about a student’s performance. Probabilistic scoring provides the mathematical mechanism to represent variation in evidence across student responses and even throughout the course of a single response. Moreover, probabilistic scoring places uncertainty directly in the scoring model of an evidence-centered design framework (Mislevy et al., 2003), allowing that uncertainty to be used directly or propagated to other inferences (e.g., those about the student).

We evaluated three zero-shot methods for scoring student writing using LLMs. While the methods showed roughly comparable accuracy across size-matched models, likelihood-based methods produced predictions with consistently superior calibration compared to judges’ self-reported confidence. Accuracy and calibration typically improve with model size. While the current study is limited to 2-way classification, other work (Christie et al., 2026, and our preliminary experiments) suggests that these findings likely extend to multi-class settings.

The methods considered here are zero-shot: they require no training data, making them flexible methods for use in assessing performance in novel tasks. However, where human-labeled examples are available, we recommend training a BERT-based classifier or fine-tuning an LLM and using either the channel or direct methods. These methods often produce substantially better accuracy than zero-shot approaches (Ferreira Mello et al., 2025), and in our preliminary experiments we also observed better ECE.

While the LLM judges we evaluated did not provide good uncertainty estimates, this is likely a property of the current generation of LLMs rather than an unavoidable limitation. Recent work demonstrates that penalizing overconfidence during post-training leads to better-calibrated LLM-produced judgments (Damani et al., 2025).

### 6.1 Limitations

All of the methods we compared are sensitive to prompt wording. Prompt sensitivity suggests a difficult optimization problem, in which tweaking or

rephrasing prompts may improve scoring performance. Automated prompt optimization is possible (Romera-Paredes et al., 2024; Wang et al., 2026), but is beyond the scope of this paper. The choice of bias-correction method also has a dramatic impact on the performance of the direct method. We report the strongest variant, but considered others (Zhao et al., 2021; Fei et al., 2023; Zhou et al., 2024); because this choice was made by comparing variants on the same evaluation set, the direct method’s reported calibration may be slightly optimistic.

While we believe that the current work provides a useful case study, we caution readers against overgeneralizing the results. We compared against only a single confidence-elicitation method for LLM judges; others are also available (Farquhar et al., 2024) and may perform differently. We also omitted any comparison to scoring from fine-tuned models or other supervised methods. Early experiments indicated that supervised approaches were substantially better on both accuracy and calibration. The current study evaluates scoring methods on a single dataset, in a single subject area. A much broader analysis is needed to establish the utility of the channel or direct methods for operational settings. Similarly, the number of models compared was small: a more thorough evaluation could include comparisons with larger open-weight models, as well as closed-weight foundation models from different model families.

Finally, while LLM likelihoods provide well-calibrated scores, our investigations also revealed a fundamental limitation: LLM likelihoods score typicality rather than correctness. In other words, a response scores highly because it meets the expectations of the prompted LLM, *not* because it is factually or logically correct. While this is an acceptable or even desirable property for some tasks, it makes likelihood-based scoring an unacceptable method for others. It also indicates a potential source of bias: surface-level dialect or stylistic features underrepresented in a model’s training data may unduly affect the scoring process, a phenomenon previously observed when using LLM judges (Ocumpaugh et al., 2025). This challenge is shared by supervised methods, but would nevertheless need to be detected and addressed prior to use with students.

## 7 Conclusion

Educators and assessors are charged with evaluating student learning against established standards such as the Common Core or state-specific variations (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010; New York State Education Department, 2017). While selected-response items are a common mechanism for such measurements, students should be able to demonstrate learning in ecologically valid ways, for example through performance tasks or authentic, motivating projects.

Generative models offer the assessment community an opportunity to build scoring and measurement processes that accumulate evidence for proficiency claims from diverse student behavior. A successful method would produce scores flexibly, against multiple claims, from observations of complex student behavior. This would enable dynamic, personalized assessments, while preserving validity and reliability. We suggest that the conditional likelihoods intrinsic to generative modeling provide an underutilized and promising avenue for producing probabilistic judgments about student learning and warrant additional investigation and validation.

## Acknowledgments

The authors thank Matthew Zent and Simon Woodhead at Eedi for valuable discussions. The authors also thank the Learning Engineering Virtual Institute, a program of Renaissance Philanthropy, for supporting this work.

## References

- S. Thomas Christie, Matthew Zent, Markus Hauru, Anna N. Rafferty, and Simon Woodhead. 2026. [Likelihood-based diagnosis with generative models: Confidence-aware measurement from student writing](#). In *Artificial Intelligence in Education: 27th International Conference, AIED 2026, Seoul, South Korea, June 27–July 3, 2026, Proceedings, Part III*, pages 236–244, Berlin, Heidelberg. Springer-Verlag.
- Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfelf, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. [Beyond binary rewards: Training LMs to reason about their uncertainty](#). Preprint, arXiv:2507.16806.
- Lawrence T DeCarlo, YoungKoung Kim, and Matthew S Johnson. 2011. A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, 48(3):333–356.
- Myroslava O 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.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nature*, 630(8017):625–630.
- Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. [Mitigating label biases for in-context learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada. 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.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. [A survey on LLM-as-a-judge](#). *The Innovation*, 7(6):101253.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.
- Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn’t always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051.
- Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-tuning: Teaching large language models to know what they don’t know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertainNLP 2024)*, pages 1–14.
- Ahmed Karim, Qiao Wang, and Zheng Yuan. 2025. Beyond the score: Uncertainty-calibrated LLMs for automated essay assessment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19642–19647.
- Mustafa Kaya and Ilyas Cicekli. 2024. A hybrid approach for automated short answer grading. *IEEE Access*, 12:96332–96341.
- Gerd Kortemeyer. 2024. [Performance of the pre-trained large language model GPT-4 on automated short answer grading](#). *Discover Artificial Intelligence*, 4(1):47.
- Shirin Lade, Alistair Willis, Jonathan Nylk, and Oli Howson. 2026. [Domain-adaptive pre-training for automated short answer grading in conceptual physics: Reliability, question-level analysis, and error reduction](#). In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pages 635–650, San Diego, California, USA. Association for Computational Linguistics.
- Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330.
- Robert J Mislevy, Linda S Steinberg, and Russell G Almond. 2003. On the structure of educational assessments. *Measurement (Mahwah NJ)*, 1(1):3–62.
- Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI’15, pages 2901–2907. AAAI Press.
- Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. [Automated evaluation of written discourse coherence using GPT-4](#). In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

- National Governors Association Center for Best Practices and Council of Chief State School Officers. 2010. Common core state standards.
- New York State Education Department. 2017. New York State P-12 Next Generation Learning Standards for Mathematics and English Language Arts. Technical report, NYSUT Research & Educational Services, Albany, NY.
- Jaclyn Ocumpaugh, Xiner Liu, and Andres Felipe Zambrano. 2025. [Language models and dialect differences](#). In *Proceedings of the 15th International Learning Analytics and Knowledge Conference, LAK '25*, pages 204–215, New York, NY, USA. Association for Computing Machinery.
- Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. 2024. [Mathematical discoveries from program search with large language models](#). *Nature*, 625(7995):468–475.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.
- Yun Wang, Xin Xia, Xuansheng Wu, Xiaoming Zhai, and Ninghao Liu. 2026. [Learnable assessment skills for LLM-based automated scoring: Rubric construction via iterative optimization](#). *Preprint*, arXiv:2605.29274.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.
- Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. 2024. [Batch calibration: Rethinking calibration for in-context learning and prompt engineering](#). In *The Twelfth International Conference on Learning Representations (ICLR)*.