Can We Trust AI's Self-assessment? Evaluating and Improving LLM Confidence Calibration in Educational Dialogue Coding

Hongming Li[1(✉)], Huan Kuang[2], and Anthony F. Botelho[1]

[1] University of Florida, Gainesville, FL 32601, USA
hli3@ufl.edu, abotelho@coe.ufl.edu
[2] Florida State University, Tallahassee, FL 32306, USA
hkuang2@fsu.edu

Abstract. Large Language Models (LLMs) are increasingly used to assist qualitative coding in educational research, but unlike traditional machine learning models that produce calibrated probabilistic outputs, LLMs report confidence as part of text generation. Whether such "verbalized" confidence can serve as a reliable indicator of prediction quality remains an open question. This study investigates LLM confidence calibration through three studies using 633 student-AI dialogues coded by experts across nine categories. First, evaluating three models with different accessibility profiles, we find universal overconfidence across all models, with discriminative ability (the capacity to rank predictions by correctness likelihood) varying substantially. Second, through an anchoring experiment, we observe that models adopt demonstrated confidence values only when they fall within a high-confidence range, completely rejecting implausibly low anchors. This asymmetric pattern, which we term plausibility-gated anchoring, suggests that verbalized confidence reflects learned priors rather than genuine uncertainty. Third, we explore activation steering as an approach to improve calibration in open-weight models by amplifying representations associated with well-calibrated predictions, finding consistent improvements that generalize to held-out data. These findings offer diagnostic insights for assessing LLM confidence reliability in coding tasks and suggest directions for improving calibration when deploying models in research workflows.

Keywords: Large Language Models · Confidence Calibration · Qualitative Coding · Human-AI Collaboration · Activation Steering · Educational Data Mining

1 Introduction

Large Language Models (LLMs) have emerged as promising tools for assisting qualitative coding in educational research, offering potential benefits in scalability, consistency, and reduced human burden [1,12,29]. However, a fundamental

E. G. Blanchard et al. (Eds.): AIED 2026, LNAI 16582, pp. 262–276, 2027.
https://doi.org/10.1007/978-3-032-29755-6_18

challenge remains: determining when AI-generated codes/labels can be trusted. Unlike traditional machine learning models that produce calibrated probabilistic outputs (i.e. [8]), LLM "confidence" is self-reported, meaning the model generates a numerical estimate of its own certainty as part of its text output [7,26]; without a mechanism to observe its own internal node activation patterns, however, an LLM is producing this confidence based solely on contextual cues as it would any set of output tokens. If such confidence scores were well-calibrated, they could enable efficient human-AI collaboration workflows where high-confidence outputs are accepted with minimal review while low-confidence outputs are flagged for human verification [16].

The reliability of these self-reported confidence scores remains largely unexplored, particularly in educational contexts where the stakes of misclassification can affect research reliability and validity. This question carries particular urgency for open-weight models, which are often the only viable option in educational settings due to constraints such as Family Educational Rights and Privacy Act (FERPA) compliance requirements that prohibit student data from leaving institutional control [19,23], budget limitations that make API costs prohibitive at scale, and reproducibility standards that demand transparent and replicable methods [22].

This study seeks to address three research questions:

- **RQ1:** How well-calibrated are LLM confidence scores across different models in educational coding tasks?
- **RQ2:** What mechanisms govern LLM confidence outputs, specifically how do demonstration examples influence self-reported confidence?
- **RQ3:** Can activation-level interventions improve confidence calibration in open-weight models?

Through three complementary studies, we provide diagnostic findings on cross-model calibration, discover a plausibility-gated anchoring mechanism that constrains prompt-based interventions, and explore activation steering as an alternative approach to improving calibration.

2 Related Work

AI-Assisted Qualitative Coding. Recent advances in natural language processing have prompted growing interest in using LLMs to support qualitative analysis in educational research [3,25]. Much of this work evaluates LLMs by their post hoc agreement with human coders, and several studies report promising performance on coding and classification tasks [5,34]; however, scholars caution that interpretability and epistemic trust remain unresolved, especially when LLM outputs are incorporated into research claims [20]. Within educational data mining and learning analytics, recent studies have moved beyond isolated benchmarks toward workflow-level investigations, showing that LLM effectiveness varies systematically with dataset and construct characteristics and with prompting choices [15], and that LLMs can also assist upstream codebook

development by leveraging established educational theories, albeit with prompt-mediated trade-offs between theoretical alignment and practical usability [33]. This literature suggests growing feasibility for LLM-assisted qualitative workflows but also reveals a key gap: existing evaluations largely treat model outputs as categorical decisions to score after the fact, offering limited insight into whether LLMs can reliably communicate their own uncertainty and thereby support principled decisions about when outputs should be trusted, deferred, or subjected to additional human review.

Confidence Calibration. Calibration refers to the alignment between predicted confidence and actual accuracy; a well-calibrated model that reports 80% confidence should be correct approximately 80% of the time [6]. Standard metrics include Expected Calibration Error (ECE) [21] and AUROC/AUC [9]. While calibration has been extensively studied in probabilistic classification models [8,18], how these notions extend to large language models remains an open question. Recent work has begun to examine confidence elicitation and self-evaluation in LLMs [24,32], but empirical evidence on whether such confidence signals are reliable or useful for downstream decision-making is still limited.

Prompt Sensitivity and Anchoring. LLM outputs are known to be sensitive to prompt formulation, including the choice of examples in few-shot settings [2,14]. Anchoring bias, the tendency for judgments to be influenced by initially presented values, is well-documented in human cognition [28] and has been observed in LLM numerical reasoning [10]. However, whether anchoring effects extend to self-reported confidence and whether such effects are symmetric or constrained by semantic plausibility remains unexplored.

Activation Steering. Recent work has demonstrated that LLM behavior can be modified by intervening on internal activations during inference [13,27]. By identifying directions in activation space associated with target behaviors such as truthfulness or refusal, researchers have shown that adding steering vectors to intermediate layer activations can shift model outputs without retraining [35]. This approach offers potential for improving calibration by steering models toward states associated with well-calibrated predictions.

3 Method

3.1 Dataset and Model Selection

We used a dataset of 633 student-AI dialogues collected from a graduate-level introductory computer science course.[1] Each dialogue consisted of a student prompt submitted to ChatGPT for assistance with course-related tasks. Two expert coders (educational researchers with training in qualitative methods) independently coded each prompt into one of nine mutually exclusive categories

[1] Data were collected with University of Florida Institutional Review Board approval (Protocol #IRB202202047).

using a coding scheme developed through iterative refinement and pilot coding by the research team in a prior study of student-AI interaction patterns. The categories are summarized in Table 1. The reliability of the coding scheme was evaluated through inter-rater reliability (IRR), with Cohen's kappa (κ) values ranging from 0.65 to 0.94 across categories, indicating substantial to strong agreement [4]. Table 1 provides the complete coding scheme with definitions and examples used for both human coding and LLM prompting.

Table 1. Coding scheme for student-AI dialogue classification with inter-rater reliability. κ values indicate Cohen's kappa between two expert coders.

Category	Definition	Examples	κ
collaboration	Student asks ChatGPT to create new content or reflects on learning	"Can you write a lesson plan?"	.88
exploration	Student raises questions expanding task scope or poses hypotheticals	"How has this topic evolved?"	.76
investigation	Student asks clarifying follow-up questions	"Could you provide more details?"	.68
resource	Student asks for external resources and references	"Are there movies on this?"	.65
framing	Student specifies audience or format for the response	"Explain for high schoolers"	.94
value judgment	Student asks questions requiring evaluative responses	"Which would you include?"	.70
default	Prompts copied directly from assignment description	"What is documentation?"	.87
example	Student asks for real-world examples or code snippets	"Give Java examples"	.84
NA	Greetings, off-topic, or empty messages	"Hello", "Thanks"	.83

We evaluated three widely-utilized LLM models representing different accessibility profiles:

- **gpt-5-mini** (OpenAI): A closed-source commercial model accessed via API, representing commercial LLMs optimized for cost-effective deployment.
- **gemini-3-flash-preview** (Google): A closed-source commercial model accessed via API, representing an alternative commercial option with competitive performance.
- **llama-3.1-8b-instruct** (Meta): An open-weight model with 8 billion parameters, deployed locally with Python using Hugging Face Transformers library [31] with 16-bit floating-point precision and automatic device mapping.

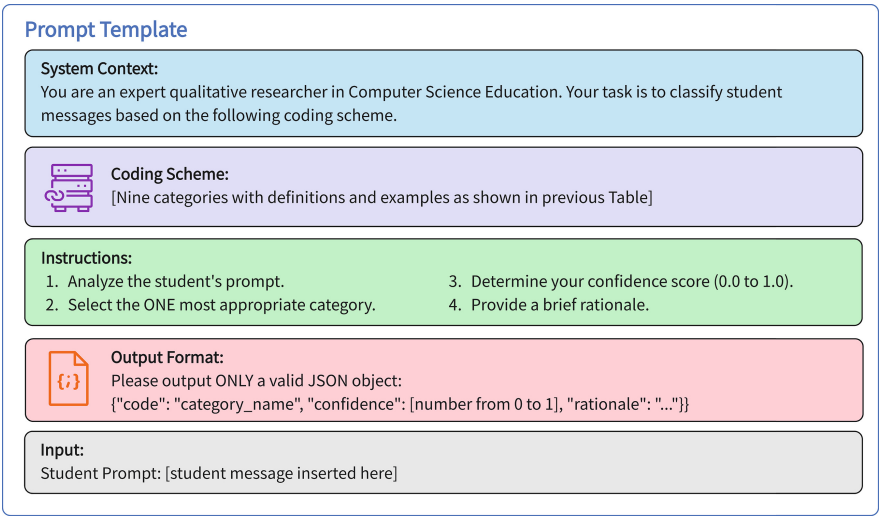


Fig. 1. Prompt template used for eliciting category predictions, confidence scores, and rationales from all evaluated models. The bracketed placeholder indicates where the coding scheme (Table 1) was inserted.

3.2 Study 1: Cross-Model Calibration Evaluation

All models received identical prompts requesting: (1) a predicted category, (2) a confidence score from 0.0 to 1.0, and (3) a brief rationale. The prompt included the complete coding scheme with definitions and examples. Figure 1 illustrates the complete prompt template used across all models. All models were run with greedy decoding (temperature = 0.0) and a maximum output length of 256 tokens to ensure deterministic, reproducible outputs. We focused on verbalized confidence, where the model self-reports a numerical score as part of its text output, rather than extracting token-level logit probabilities. This choice reflects the dominant paradigm in applied LLM workflows where practitioners elicit confidence through prompting, and ensures comparability across closed-source and open-weight models, as logit access is unavailable for commercial APIs.

Following established practices in calibration research [8], we evaluated calibration using multiple complementary metrics. *Accuracy* measured agreement with expert codes. *AUC* assessed the ability of confidence to discriminate correct from incorrect predictions, where 0.5 indicates random performance and 1.0 indicates perfect discrimination [9]. Following conventions in the calibration literature, AUC values below 0.6 suggest limited practical utility for distinguishing correct from incorrect predictions, values between 0.6 and 0.7 indicate moderate discriminative ability, and values above 0.7 are generally considered adequate for confidence-based decision-making [17]. *Spearman's* ρ captured rank correlation between confidence and correctness. *Expected Calibration Error* (ECE), com-

puted with 10 bins, measured the average gap between confidence and accuracy, where lower values indicate better calibration [21].

3.3 Study 2: Anchoring Effect Analysis

Studies 2 and 3 focused on the open-weight llama-3.1-8b-instruct model for three reasons. First, open-weight models are often the only viable option in educational settings with data governance constraints, making their calibration properties particularly consequential for practice. Second, as is reported later in Sect. 4.1, the open-weight model exhibited the weakest calibration in Study 1, suggesting both greater need and greater opportunity for improvement. Finally, and importantly, Study 3 requires access to internal model activations, which is only possible with open-weight models (discussed in the next section). To investigate how demonstration examples influence confidence outputs, we manipulated the confidence value shown in the output format specification within the prompt. Three conditions were tested:

- **Baseline**: The format example showed "confidence: [number from 0 to 1]" with no specific value.
- **High Anchor (0.95)**: The format example showed "confidence: 0.95".
- **Low Anchor (0.05)**: The format example showed "confidence: 0.05".

All other prompt components remained identical across conditions. We analyzed the confidence distributions using descriptive statistics (mean, standard deviation, interquartile range, kurtosis) and computed the proportion of outputs falling within 0.02 of each anchor value to quantify anchor adoption rates.

3.4 Study 3: Activation Steering Intervention

To explore whether activation-level interventions could improve calibration in the open-weight model, we implemented an activation steering approach following recent work on representation engineering [27,35]. First, we extracted activations from an intermediate layer (layers 16 and 22 were tested, representing middle and upper-middle positions in the 32-layer architecture) during inference on all 633 samples. We then partitioned samples based on calibration quality using a confidence threshold of 0.90: *well-calibrated* samples were those where confidence and correctness aligned (high confidence and correct, or low confidence and incorrect), while *poorly-calibrated* samples showed misalignment (high confidence but incorrect, or low confidence but correct). A steering vector was computed as the difference between mean activations of well-calibrated and poorly-calibrated samples. During subsequent inference, this vector was added to the target layer's activations, scaled by a coefficient. We tested two configurations pairing layer position with coefficient magnitude: a middle layer (16) with a moderate coefficient (2.0) and an upper-middle layer (22) with a larger coefficient (3.0), as higher layers encode more task-specific representations that may require stronger amplification to influence output behavior.

Table 2. Cross-model calibration comparison on 633 samples. ECE = Expected Calibration Error (lower is better). AUC and Spearman's ρ measure discriminative ability (higher is better).

Model	Accuracy	Mean Conf.	ECE	AUC	Spearman ρ
gpt-5-mini	0.441	0.908	0.467	0.689	0.331
gemini-3-flash-preview	0.520	0.922	0.403	0.669	0.310
llama-3.1-8b-instruct	0.338	0.891	0.553	0.565	0.131

To ensure valid evaluation and prevent information leakage [11], we employed a 50/50 train-test split with a fixed random seed for reproducibility: the steering vector was computed using only the training set (316 samples), while generalization was assessed on the held-out test set (317 samples). Within the training set, 128 samples were classified as well-calibrated and 188 as poorly-calibrated, reflecting more overconfident, incorrect predictions among the baseline model.

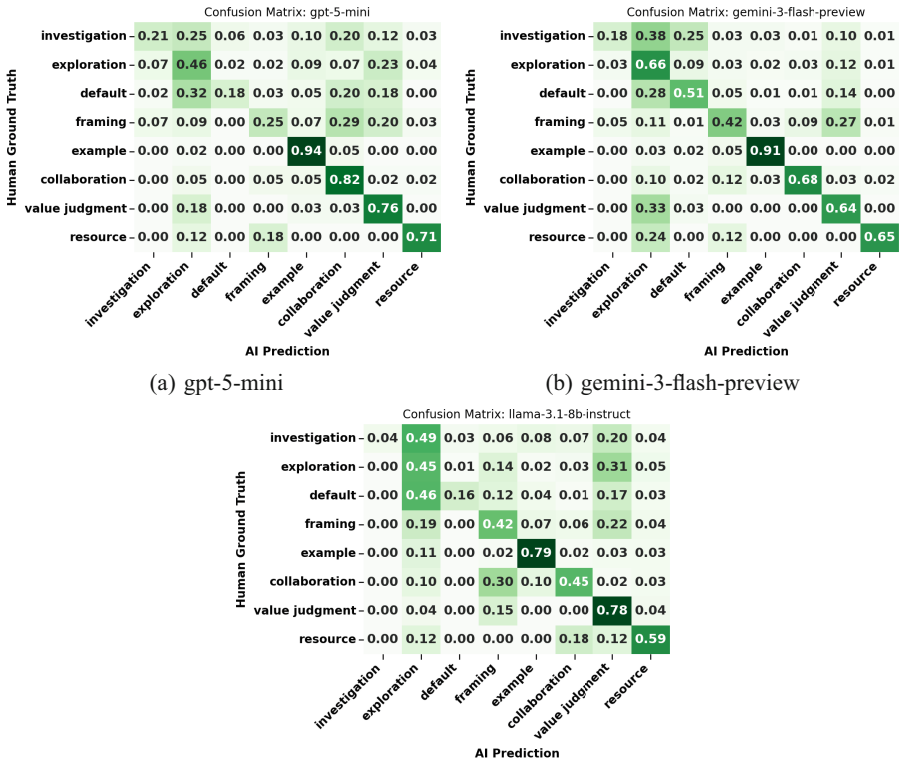
4 Results

4.1 Study 1: Cross-Model Calibration

Table 2 presents calibration metrics across all three models. All models exhibited substantial overconfidence: mean confidence scores ranged from 0.891 to 0.922, while actual accuracy ranged from 0.338 to 0.520, yielding confidence-accuracy gaps of 0.40 to 0.55. These accuracy levels reflect the inherent difficulty of nine-way classification (chance baseline ≈ 0.11) across semantically overlapping categories, consistent with prior findings that LLM coding performance varies substantially with task granularity and construct complexity [15]. The ECE values (0.403 to 0.553) confirm that stated confidence systematically exceeded realized accuracy across all models. Despite this shared tendency toward overconfidence, discriminative ability varied substantially. The closed-source models achieved AUC values of 0.689 (gpt-5-mini) and 0.669 (gemini-3-flash-preview), indicating moderate ability to rank predictions by likelihood of correctness, while the open-weight llama-3.1-8b-instruct achieved only 0.565, marginally above the 0.5 baseline representing random discrimination. This pattern was mirrored in Spearman correlations: 0.331 and 0.310 for closed-source models versus 0.131 for the open-weight model. Notably, accuracy alone did not determine calibration quality: gemini-3-flash-preview achieved the highest accuracy (0.520) but slightly lower discriminative ability (AUC 0.669) than gpt-5-mini (accuracy 0.441, AUC 0.689). Figure 2 presents confusion matrices illustrating classification error patterns across the nine coding categories.

4.2 Study 2: Anchoring Effect Analysis

Table 3 presents confidence distribution metrics under different anchoring conditions. The high anchor condition (0.95) produced dramatic mode collapse: 96.5%



(a) gpt-5-mini

(b) gemini-3-flash-preview

(c) llama-3.1-8b-instruct

Fig. 2. Normalized confusion matrices for each model. Rows represent ground truth categories; columns represent model predictions. All models show concentration of predictions in a subset of categories, with varying error patterns across models.

of outputs fell within 0.02 of the anchor value, standard deviation dropped from 0.055 to 0.014, and kurtosis increased to 37.72, indicating an extremely peaked distribution. In stark contrast, the low anchor condition (0.05) produced zero adoption; no outputs fell within 0.02 of the anchor. Instead, the distribution reverted toward the baseline high-confidence regime, with mean confidence (0.884) remaining close to baseline (0.891) and similar standard deviation (0.058 vs. 0.055). The negative kurtosis (−1.11) and increased IQR (0.150) indicate a flatter distribution, but one still concentrated in the high-confidence range.

The discrete value distribution (Table 4) further illuminates this pattern. At baseline, 99% of outputs fell on just three values: 0.90 (68.2%), 0.80 (20.1%), and 1.00 (10.7%). Under the low anchor condition, outputs redistributed among these same high-confidence values rather than shifting toward 0.05, with the distribution spreading to include 0.95 (24.0%) alongside 0.90 (41.2%) and 0.80 (27.0%). Figure 3 visualizes these distributions, showing that high anchoring

induces mode collapse, while low anchoring causes predictions to revert to the model's default, high-confidence prediction distribution.

Table 5 shows how anchoring affected calibration metrics. The high anchor condition degraded discriminative ability: AUC dropped from 0.565 to 0.518 and Spearman's ρ from 0.131 to 0.095, as variance collapse eliminated the model's capacity to differentiate correct from incorrect predictions. The low anchor condition produced metrics similar to baseline (AUC 0.564, Spearman's ρ 0.111), consistent with the distribution reverting to its default state.

Table 3. Confidence distribution metrics under different anchoring conditions for llama-3.1-8b-instruct. P(≈0.95) = proportion of outputs within 0.02 of 0.95. P(≈0.05) = proportion within 0.02 of 0.05.

Condition	Mean	SD	IQR	Kurtosis	P(≈0.95)	P(≈0.05)
Baseline	0.891	0.055	0.000	0.18	0.9%	0.0%
Anchor = 0.95	0.949	0.014	0.000	37.72	96.5%	0.0%
Anchor = 0.05	0.884	0.058	0.150	−1.11	24.0%	0.0%

Table 4. Three most frequent confidence values under each condition for llama-3.1-8b-instruct.

Condition	Rank 1	Rank 2	Rank 3
Baseline	0.90 (68.2%)	0.80 (20.1%)	1.00 (10.7%)
Anchor = 0.95	0.95 (96.5%)	1.00 (1.4%)	0.85 (1.4%)
Anchor = 0.05	0.90 (41.2%)	0.80 (27.0%)	0.95 (24.0%)

4.3 Study 3: Activation Steering

Table 6 presents the effects of activation steering on calibration metrics, evaluated on the held-out test set not used to compute the steering vector. Both configurations produced consistent improvements: AUC increased from 0.565 to 0.588 (layer 16, coefficient 2.0) and 0.584 (layer 22, coefficient 3.0), representing absolute gains of 0.019 to 0.023. The relative improvements in Spearman's ρ were more substantial, reaching 29.0% for layer 16 and 23.7% for layer 22, indicating that steering improved the rank ordering of confidence scores relative to correctness. While modest in absolute magnitude, these improvements were consistent across both layer positions and steering coefficients tested, and the fact that they generalized to held-out data suggests the steering vector captures a genuine pattern in activation space rather than overfitting to training samples. The absolute calibration gap remained large, indicating that activation steering partially but not fully addresses the overconfidence.

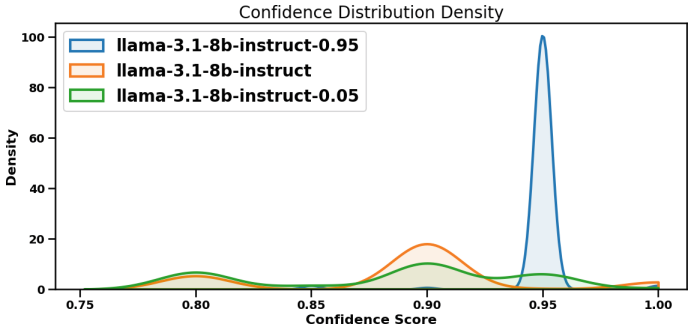


Fig. 3. Confidence distributions under three anchoring conditions for llama-3.1-8b-instruct. The high anchor (0.95) produces mode collapse with 96.5% of outputs at exactly 0.95. The low anchor (0.05) is rejected, and the distribution reverts to the high-confidence prior with outputs around 0.80, 0.90, and 0.95.

Table 5. Calibration metrics under different anchoring conditions for llama-3.1-8b-instruct.

Condition	Accuracy	AUC	Spearman ρ
Baseline	0.338	0.565	0.131
Anchor = 0.95	0.344	0.518	0.095
Anchor = 0.05	0.332	0.564	0.111

5 Discussion

Synthesis of Findings. The three studies form a coherent progression. Study 1 (Table 2) established universal overconfidence across model families, with discriminative ability varying substantially: closed-source models showed moderate utility for ranking predictions (AUC in the 0.67 to 0.69 range), while the open-weight model performed only marginally above chance. Study 2 (Tables 3 and 4) then revealed a mechanism underlying the open-weight model's limitation: confidence outputs concentrated on a discrete set of high values, and the model adopted demonstrated anchors only when they fell within this high-confidence range. Study 3 (Table 6) demonstrated that activation steering could partially bypass this constraint, with improvements generalizing to held-out data. These findings suggest that verbalized LLM confidence reflects learned semantic priors rather than probabilistic uncertainty estimation in the traditional sense [6,8].

Discriminative Ability Beyond Accuracy. A notable finding from Study 1 is that classification accuracy did not determine calibration quality. The model with the highest accuracy (gemini-3-flash-preview, 0.520) showed slightly lower discriminative ability (AUC 0.669) than a less accurate model (gpt-5-mini, 0.441 accuracy, AUC 0.689). This dissociation suggests that confidence calibration depends not merely on whether the model "knows" the correct answer, but on

whether the model's internal uncertainty tracking aligns with its error patterns. A model may achieve high accuracy through consistent performance on common categories while remaining poorly calibrated on edge cases; conversely, a model with lower accuracy might produce confidence scores that more faithfully track its own limitations. This finding reinforces that calibration should be evaluated as a distinct property rather than assumed to follow task performance [8,18].

Table 6. Activation steering hold-out set results on llama-3.1-8b-instruct. Δ columns show change from baseline. L = layer, C = steering coefficient.

Condition	AUC	Δ AUC	Δ%	Spearman ρ	$\Delta\rho$	Δ%
Baseline	0.565	–	–	0.131	–	–
Steered (L16, C=2.0)	0.588	+0.023	+4.1%	0.169	+0.038	+29.0%
Steered (L22, C=3.0)	0.584	+0.019	+3.4%	0.162	+0.031	+23.7%

Plausibility-Gated Anchoring. We term the asymmetric pattern observed in Study 2 *plausibility-gated anchoring*: LLMs adopt demonstrated numerical values only when those values fall within a range the model treats as semantically plausible for the task type. This phenomenon extends prior work on anchoring bias in human cognition [28] and LLM numerical reasoning [10], adding an important constraint: anchoring is not symmetric but filtered through learned priors. One plausible explanation is that the model's training data contains abundant examples of classification tasks accompanied by high reported confidence, establishing a prior that treats low confidence as atypical for such tasks. The discrete structure of baseline outputs supports this interpretation: 99% of values fell on just three points (0.80, 0.90, 1.00), and exposure to the low anchor (0.05) caused redistribution among these same high values rather than a shift toward the anchor. For researchers seeking to improve calibration through prompting, this finding identifies one important constraint: demonstrating lower confidence values through format examples alone did not overcome the model's intrinsic prior in our experiments. Other prompting strategies, such as providing multiple examples with varied confidence levels or explicitly describing conditions that warrant lower confidence, may prove more effective and warrant future investigation.

Activation-Level Intervention. Given Study 2's finding that prompt-level interventions are constrained by the model's plausibility prior, Study 3 investigated whether activation-level interventions could bypass this limitation. The results demonstrate that it can: steering vectors computed from the difference between well-calibrated and poorly-calibrated samples produced consistent improvements across both layer positions (16 and 22) and steering coefficients (2.0 and 3.0) tested. Crucially, these improvements generalized to held-out data not used in computing the steering vector, indicating that the identified direction captures a genuine pattern in activation space rather than overfitting to training samples. The improvement in Spearman's ρ (+29%) is particularly notable,

as it reflects enhanced rank ordering of confidence relative to correctness. This finding establishes that calibration-relevant information exists within model representations and can be amplified through targeted intervention, offering a complementary approach to prompt-based strategies. Unlike post-hoc recalibration methods such as temperature scaling [8] that operate on model outputs, activation steering intervenes during the inference process itself, suggesting these approaches may be combined in future work for greater effect.

Implications for Human-AI Collaboration. For educational researchers designing coding workflows, our findings indicate that the viability of confidence-based triage depends on discriminative ability. An AUC of 0.68, as observed with the closed-source models, means that a randomly selected correct prediction will receive higher confidence than a randomly selected incorrect prediction 68% of the time. This level of discrimination, while far from perfect, can meaningfully prioritize items for human review by directing attention toward lower-confidence predictions [16]. In contrast, an AUC near 0.56, as observed with the open-weight model, provides only marginally better than random ranking, limiting the practical value of confidence-based filtering. Critically, even when discriminative ability is adequate, the substantial overconfidence (ECE > 0.40) means raw confidence values should not be interpreted as probability estimates; they serve better as relative rankings than as absolute indicators of correctness likelihood. This distinction matters particularly for educational settings subject to data governance requirements [23], where local deployment of open-weight models may be the only viable option; the activation steering approach demonstrated in Study 3 offers one path toward improving reliability in such cases, though it requires additional technical setup beyond standard prompting. To support adoption, we release our activation steering implementation as an open-source toolkit.[2] Regardless of deployment context, our findings suggest that including explicit confidence values in few-shot examples risks inducing mode collapse or may otherwise be ignored entirely, depending on whether the demonstrated value falls within the model's plausibility range.

Limitations and Future Directions. Several factors constrain generalizability. First, we evaluated a single dataset from introductory computer science with a nine-category coding scheme; replication across educational domains and coding schemes of varying granularity would strengthen these conclusions. Second, our open-weight models evaluation focused on llama-3.1-8b-instruct; larger models with 70 billion or more parameters may exhibit different calibration properties, and the boundaries of plausibility priors may shift with scale. Third, while the activation steering improvements generalized to held-out data within our dataset, evaluation across additional datasets would further establish robustness. Fourth, we did not examine per-category calibration, though constructs like *default* may be inherently easier to recognize than *collaboration*, yielding category-specific confidence reliability; future work reporting category-level met-

[2] Source code available at https://github.com/hichipli/LLM-confidence-calibration-in-educational-dialogue-coding-AIED2026.

rics would provide more actionable guidance for researchers deciding which coding categories can be safely automated versus those requiring human review. Finally, our evaluation used single-turn, zero-shot prompting; multi-turn interactions, chain-of-thought prompting [30], or explicit uncertainty elicitation may alter calibration properties.

6 Conclusion

This work contributes to the literature on AI-assisted qualitative research in education [3,20,25] in three ways. First, we provide diagnostic evidence that confidence calibration merits scrutiny before deploying LLMs in coding workflows, documenting universal overconfidence alongside meaningful variation in discriminative ability across model families. Second, we identify plausibility-gated anchoring as a mechanism constraining prompt-based calibration interventions, explaining why verbalized confidence reflects learned priors rather than epistemic uncertainty and why simply demonstrating lower confidence values cannot overcome this constraint. Third, we demonstrate that activation steering can improve calibration in open-weight models, with improvements that generalize to held-out data and thus reflect genuine patterns rather than artifacts of the training samples. These findings advance understanding of when LLM confidence can support human-AI collaboration in educational coding, providing diagnostic tools for evaluating model trustworthiness and an intervention approach that complements prompt-based strategies.

Acknowledgments. We would like to thank the National Science Foundation (#2331379), Institute of Education Sciences (#R305B230007), Gates Foundation (#078981), Support from the Learning Engineering Tools Competition, and other anonymous philanthropy.

References

1. Borchers, C., Yang, K., Lin, J., Rummel, N., Koedinger, K.R., Aleven, V.: Combining dialog acts and skill modeling: what chat interactions enhance learning rates during AI-supported peer tutoring? In: Proceedings of the 17th International Conference on Educational Data Mining, pp. 117–130 (2024)
2. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
3. Chew, R., Bollenbacher, J., Wenger, M., Speer, J., Kim, A.: LLM-assisted content analysis: using large language models to support deductive coding. arXiv preprint arXiv:2306.14924 (2023)
4. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
5. Dai, S.C., Xiong, A., Ku, L.W.: LLM-in-the-loop: leveraging large language model for thematic analysis. arXiv preprint arXiv:2310.15100 (2023)
6. DeGroot, M.H., Fienberg, S.E.: The comparison and evaluation of forecasters. J. Roy. Stat. Soc.: Ser. D (The Statistician) **32**(1–2), 12–22 (1983)

7. Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., Gurevych, I.: A survey of confidence estimation and calibration in large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6577–6595 (2024)
8. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR (2017)
9. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
10. Jones, E., Steinhardt, J.: Capturing failures of large language models via human cognitive biases. Adv. Neural. Inf. Process. Syst. **35**, 11785–11799 (2022)
11. Kapoor, S., Narayanan, A.: Leakage and the reproducibility crisis in ml-based science (2022). https://arxiv.org/abs/2207.07048
12. Kasneci, E., et al.: ChatGPT for good? On opportunities and challenges of large language models for education. Learn. Individ. Differ. **103**, 102274 (2023)
13. Li, K., Patel, O., Viégas, F., Pfister, H., Wattenberg, M.: Inference-time intervention: eliciting truthful answers from a language model. Adv. Neural. Inf. Process. Syst. **36**, 41451–41530 (2023)
14. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., Neubig, G.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 1–35 (2023)
15. Liu, X., et al.: Qualitative coding with GPT-4: where it works better. J. Learn. Anal. **12**(1), 169–185 (2025)
16. Lubars, B., Tan, C.: Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
17. Mandrekar, J.N.: Receiver operating characteristic curve in diagnostic test assessment. J. Thorac. Oncol. **5**(9), 1315–1316 (2010)
18. Minderer, M., et al.: Revisiting the calibration of modern neural networks. Adv. Neural. Inf. Process. Syst. **34**, 15682–15694 (2021)
19. Mithun, P., Noriega-Atala, E., Merchant, N., Skidmore, E.: A gateway for egalitarian access to LLM based resources. In: Cristea, A.I., Walker, E., Lu, Y., Santos, O.C., Isotani, S. (eds.) AIED 2025. CCIS, vol. 2590, pp. 137–146. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-99261-2_13
20. Naeem, H., Hauser, J.: Should we discourage AI extension? Epistemic responsibility and AI. Philos. Technol. **37**(3), 91 (2024)
21. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using Bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 29 (2015)
22. National Academies of Sciences and Medicine and Policy and Global Affairs and Board on Research Data and Division on Engineering and Physical Sciences and Committee on Applied and Theoretical Statistics and Board on Mathematical Sciences and others: Reproducibility and replicability in science. National Academies Press (2019)
23. Reidenberg, J.R., Schaub, F.: Achieving big data privacy in education. Theory Res. Educ. **16**(3), 263–279 (2018)
24. Ren, J., Zhao, Y., Vu, T., Liu, P.J., Lakshminarayanan, B.: Self-evaluation improves selective generation in large language models. In: Proceedings on. pp. 49–64. PMLR (2023)

25. Tai, R.H., et al.: An examination of the use of large language models to aid analysis of textual data. Int. J. Qualit. Meth. **23**, 16094069241231168 (2024)
26. Tian, K., et al.: Just ask for calibration: strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv preprint arXiv:2305.14975 (2023)
27. Turner, A.M., et al.: Steering language models with activation engineering. arXiv preprint arXiv:2308.10248 (2023)
28. Tversky, A., Kahneman, D.: Judgment under uncertainty: heuristics and biases: biases in judgments reveal some heuristics of thinking under uncertainty. Science **185**(4157), 1124–1131 (1974)
29. Venugopalan, D., Yan, Z., Borchers, C., Lin, J., Aleven, V.: Combining large language models with tutoring system intelligence: a case study in caregiver homework support. In: Proceedings of the 15th Int'l Learning Analytics and Knowledge Conference, pp. 373–383 (2025)
30. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural. Inf. Process. Syst. **35**, 24824–24837 (2022)
31. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
32. Yang, D., Tsai, Y.H.H., Yamada, M.: On verbalized confidence scores for LLMs. arXiv preprint arXiv:2412.14737 (2024)
33. Zambrano, A.F., et al.: Data plus theory equals codebook: leveraging LLMs for human-AI codebook development. J. Educ. Data Min. **18**(1), 25–65 (2026)
34. Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., Yang, D.: Can large language models transform computational social science? Comput. Linguist. **50**(1), 237–291 (2024)
35. Zou, A., et al.: Representation engineering: a top-down approach to AI transparency. arXiv preprint arXiv:2310.01405 (2023)