https://doi.org/10.1038/s44355-026-00053-3

Across generations, sizes, and types, large language models poorly report self-confidence in gastroenterology clinical reasoning tasks

Check for updates

Nariman Naderi[1,7], Seyed Amir Ahmad Safavi-Naini[1,2,7], Thomas Savage[3], Mohammad Amin Khalafi[4], Peter R. Lewis[5], Zahra Atf[5], Girish Nadkarni[1,2] & Ali Soroush[1,2,6] ✉

This study evaluated confidence calibration across 48 large language models (LLMs) using 300 gastroenterology board exam-style multiple-choice questions. Regardless of accuracy, all models demonstrated poor self-estimation of certainty. Even the best-calibrated systems (o1 preview, GPT-4o, Claude-3.5-Sonnet) exhibited substantial overconfidence (Brier scores 0.15–0.2, AUROC ~ 0.6). Models maintained high confidence regardless of question difficulty or response correctness. In their current form, LLMs cannot be relied upon to communicate uncertainty, and human oversight remains essential for safe use.

Reliable communication of uncertainty is essential for safe clinical care. Large Language Models (LLMs) can generate fluent, clinically relevant answers, but they can also present incorrect information with unwarranted certainty[1,2]. In patient care, misplaced trust in such outputs can lead to dangerous outcomes[3,4]. To be safe for clinical use, an LLM must not only provide accurate answers but also convey confidence levels that reflect the likelihood of being correct. Without this calibration, even highly accurate models can pose risks in clinical contexts.

Several techniques can estimate uncertainty in LLM outputs, but most are difficult to apply in routine clinical practice. They require access to internal model computations or involve complex procedures such as generating and comparing multiple responses, training additional models, or applying statistical calibration after the fact[5–7]. In addition to being unproven, these approaches demand significant computing resources, technical expertise, and interpretation skills that are not widely available in healthcare settings. As a result, there is a need for uncertainty estimation strategies that are both reliable and practical for clinicians.

Self-reported confidence is one such practical alternative. In this approach, the model is asked how certain it is about its answer, often on a simple numerical scale[8–11]. It is intuitive, easy for clinicians to interpret alongside the clinical information already provided, and requires no special access to the model's internal workings or substantial computing power. However, because the confidence statement is itself generated language, it presents an additional opportunity for an incorrect output. Its value therefore depends on two related properties: calibration, the agreement between the stated confidence and the actual probability of being correct, and discrimination, the tendency to assign higher confidence to correct than to incorrect answers. Both properties reflect an aspect of language-based metacognition, or the ability to recognize and communicate the limits of one's own knowledge. If this capability is weak or absent, verbalized confidence may mislead rather than guide clinical decision-making. Although previous studies have found evidence of poor calibration in general medical tasks[12,13], it remains unclear whether this limitation is confined to certain systems or represents a universal property of current LLMs.

We addressed this question with a systematic, cross-model evaluation in a high-stakes clinical subspecialty. We evaluated self-reported confidence for 48 commercial and open-source LLMs across local, web, and API-based environments using 300 multiple-choice, board-style questions from the 2022 American College of Gastroenterology self-assessment examination. Gastroenterology was selected as our test domain primarily because the senior author (AS) is a practicing gastroenterologist, providing direct clinical insight into the impact of LLM confidence miscalibration.

[1] Division of Data-Driven and Digital Medicine (D3M), Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2] The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3] Division of Hospital Medicine, University of Pennsylvania, Philadelphia, USA. [4] Shahid Beheshti University of Medical Sciences, Research Institute for Gastroenterology and Liver Diseases, Tehran, Iran. [5] Faculty of Business and Information Technology, Ontario Tech University, Oshawa, Canada. [6] Henry D. Janowitz Division of Gastroenterology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7] These authors contributed equally: Nariman Naderi, Seyed Amir Ahmad Safavi-Naini.
✉e-mail: ali.soroush@mountsinai.org

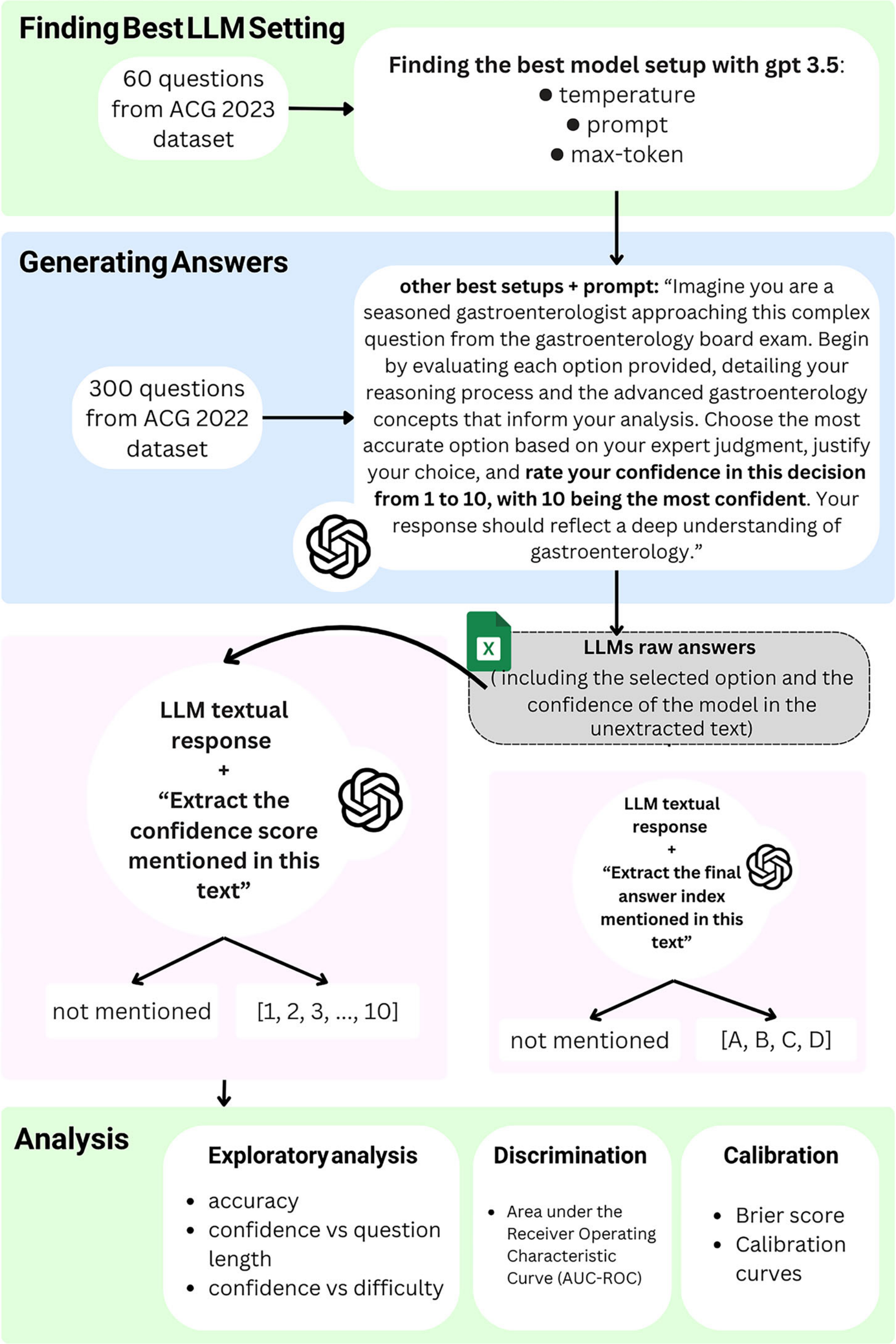
Finding Best LLM Setting
60 questions from ACG 2023 dataset
Finding the best model setup with gpt 3.5:
● temperature
● prompt
● max-token
Generating Answers
300 questions from ACG 2022 dataset
other best setups + prompt: "Imagine you are a seasoned gastroenterologist approaching this complex question from the gastroenterology board exam. Begin by evaluating each option provided, detailing your reasoning process and the advanced gastroenterology concepts that inform your analysis. Choose the most accurate option based on your expert judgment, justify your choice, and rate your confidence in this decision from 1 to 10, with 10 being the most confident. Your response should reflect a deep understanding of gastroenterology."
X
LLMs raw answers
(including the selected option and the confidence of the model in the unextracted text)
LLM textual response + "Extract the confidence score mentioned in this text"
not mentioned
[1, 2, 3, ..., 10]
LLM textual response + "Extract the final answer index mentioned in this text"
not mentioned
[A, B, C, D]
Analysis
Exploratory analysis
• accuracy
• confidence vs question length
• confidence vs difficulty
Discrimination
• Area under the Receiver Operating Characteristic Curve (AUC-ROC)
Calibration
• Brier score
• Calibration curves

Fig. 1 | Study workflow for eliciting and extracting self-reported confidence from large language models. Schematic overview of the pipeline used to present 300 gastroenterology board exam-style multiple-choice questions to 48 LLMs, record their selected answers and self-reported confidence scores on a 0—10 scale, and process raw text outputs through a structured parsing pipeline. The pipeline combines rule-based and LLM-based extraction to identify sentences containing confidence statements, extract numeric scores or flag missing confidence, and produce a curated dataset of question–answer–confidence triplets for downstream analyses.

Subspecialty domains like gastroenterology also present unique challenges for clinical reasoning that make them ideal test environments since they require integration of diverse knowledge sources to formulate diagnoses and involve procedures with significant risks, where diagnostic or treatment errors can lead to serious patient harm. We used standardized board exam-style questions because they offer an objective benchmark for evaluating model performance across a range of clinically relevant scenarios.

We employed a systematic approach where models were instructed to select the correct answer choice to each board exam question and explicitly report their confidence on a 0–10 scale (from least to most confident). Using our established methodology, we optimized model parameters including prompt instructions (the specific phrasing of text input), temperature settings (degree of model output randomness), and token limit (maximum allowable text received and generated by the model in a single response) to maximize response accuracy[14]. A semi-automated extraction pipeline with human verification (99% accuracy, Supplementary Fig. S1) was used to process the responses and confidence scores for subsequent analysis.

We extracted 13,362 answers and 12,307 confidence scores (Fig. 1). The difference between these counts resulted primarily from non-compliance with prompt instructions ($n = 846$) or from reasoning models that exhausted their token limits because of their internal reasoning dialogues ($n = 209$) (Supplementary Fig. S2). Mean confidence scores ranged from 7.99 (95% CI: 7.89–8.09) for Claude-3-Opus to 9.58 (95% CI: 9.45–9.71) for Mistral-7b, while accuracy varied substantially from 30.3% (Llama3-8b-Q8) to 81.5% (o1 preview) (Table 1). All models demonstrated

Table 1 | LLM accuracy, discrimination, calibration, and confidence scores, sorted from best calibration (lowest Brier score) to worst for each model family

Model family	Model name and parameter (quantization)	Date accessed	Calibration		Discrimination	Accuracy	Self-reported confidence score
			Brier score	ECE	AUROC	Percent	Mean (95CI)
Llama							
	Llama-3.3-70b	December 2024	0.260	0.199	0.563	65.66	8.46 (8.36–8.56)
	Llama 3.1 405B	August 2024	0.273	0.211	0.592	64	8.47 (8.38–8.57)
	Llama3.2-90B	December 2024	0.302	0.269	0.600	60.00	8.49 (8.34–8.62)
	Llama 3.1 70B	August 2024	0.313	0.283	0.538	58.19	8.51 (8.39–8.62)
	Llama 3 70B	May 2024	0.334	0.301	0.572	54.66	8.38 (8.28–8.48)
	Llama 3 8B	May 2024	0.422	0.450	0.478	43.33	8.54 (8.41–8.68)
	Llama-3.2-11b	December 2024	0.400	0.390	0.519	48.65	8.59 (8.46–8.69)
	Llama 3.1 8B	August 2024	0.433	0.441	0.512	43.14	8.67 (8.54–8.80)
	Llama-3.2-3b	December 2024	0.465	0.487	0.534	35.66	8.32 (8.18–8.45)
	Llama 2 70B	April 2024	0.481	0.493	0.529	37.71	8.70 (8.58–8.81)
	Llama-3.2-1b	December 2024	0.500	0.511	0.455	30.61	8.13 (7.96–8.31)
	Llama 2 13B (Q5)	April 2024	0.525	0.546	0.5	35.16	8.98 (8.92–9.04)
	Llama 3 8B (Q8)	April 2024	0.527	0.613	0.472	30.35	8.65 (8.28–9.02)
	Llama 2 7B	April 2024	0.528	0.587	0.47	30.87	8.66 (8.47–8.84)
	Llama 2 13B	April 2024	0.531	0.558	0.52	33.11	8.89 (8.82–8.95)
	Llama 2 7B (Q8)	April 2024	0.559	0.582	0.458	32.45	9.07 (8.98–9.15)
Qwen							
	Qwen-2.5-72b	September 2024	0.326	0.304	0.549	61.48	8.39(8.15–8.63)
	Qwen-2-72B	September 2024	0.364	0.360	0.583	57.00	9.10(8.98–9.20)
Phi							
	Phi-3 Medium 14B (Q6)	April 2024	0.389	0.377	0.588	48.66	8.57 (8.48–8.67)
	Phi-3 3B FP16	April 2024	0.458	0.464	0.486	43.79	8.96 (8.84–9.07)
	Phi-3.5-4b	December 2024	0.558	0.578	0.465	31.86	8.96 (8.90–9.02)
Google							
	Gemini Advanced Web	March–April 2024	0.297	0.247	0.561	58.49	8.20 (8.07–8.33)
	Gemma 2 27B	July 2024	0.374	0.352	0.557	50	8.52 (8.41–8.63)
	Gemma 2 9B (Q8)	July 2024	0.397	0.392	0.543	45.33	8.40 (8.30-8.50)
	Gemma 2 9B	July 2024	0.398	0.390	0.592	44.59	8.33 (8.20–8.45)
	Gemini Web	March 2024	0.421	0.420	0.563	44.44	8.61 (8.53–8.70)
Mistral							
	Mistral Large	April 2024	0.282	0.224	0.602	60.53	8.13 (7.98–8.28)
	Mixtral 8x7B	April 2024	0.359	0.336	0.547	54.33	8.79 (8.72–8.87)
	Mistral v2 Q8	April 2024	0.506	0.527	0.554	39.06	9.11 (8.90–9.32)
	Mistral 7B	April 2024	0.547	0.551	0.519	40.66	
Claude							
	Claude 3.5 Sonnet	July 2024	0.207	0.122	0.6	74	8.60 (8.54–8.67)
	Claude 3 Opus	March–April 2024	0.229	0.150	0.575	70.35	8.54 (8.44–8.63)
	Claude 3 Opus Web	March–April 2024	0.246	0.154	0.578	65.66	7.99 (7.89–8.09)
	Claude 3 Sonnet Web	March–April 2024	0.326	0.284	0.551	55.33	8.37 (8.29–8.45)
	Claude 3 Sonnet	March–April 2024	0.361	0.336	0.559	51.17	8.48 (8.39–8.58)
	Claude 3 Haiku	March–April 2024	0.373	0.357	0.522	53.76	8.88 (8.80–8.96)
	Claude 3 Haiku Web	March–April 2024	0.398	0.385	0.523	50	8.85 (8.80–8.90)
GPT							
	o1 preview	September 2024	0.157	0.100	0.576	81.57	9.15 (9.10–9.20)
	GPT-4o	May 2024	0.208	0.148	0.604	74	8.86 (8.80–8.92)
	GPT-4 Web	March 2024	0.267	0.221	0.588	66.22	8.79 (8.70–8.87)
	GPT-4	March 2024	0.278	0.237	0.605	66.53	9.02 (8.92–9.13)

Table 1 (continued) | LLM accuracy, discrimination, calibration, and confidence scores, sorted from best calibration (lowest Brier score) to worst for each model family

Model family	Model name and parameter (quantization)	Date accessed	Calibration		Discrimination	Accuracy	Self-reported confidence score
			Brier score	ECE	AUROC	Percent	Mean (95CI)
	o1 Mini	September 2024	0.278	0.257	0.626	66.33	9.20 (9.12–9.27)
	GPT-4o Mini	July 2024	0.342	0.309	0.572	56.61	8.75 (8.67–8.83)
	GPT-3.5 Web	March 2024	0.394	0.375	0.546	47.66	8.56 (8.48–8.63)

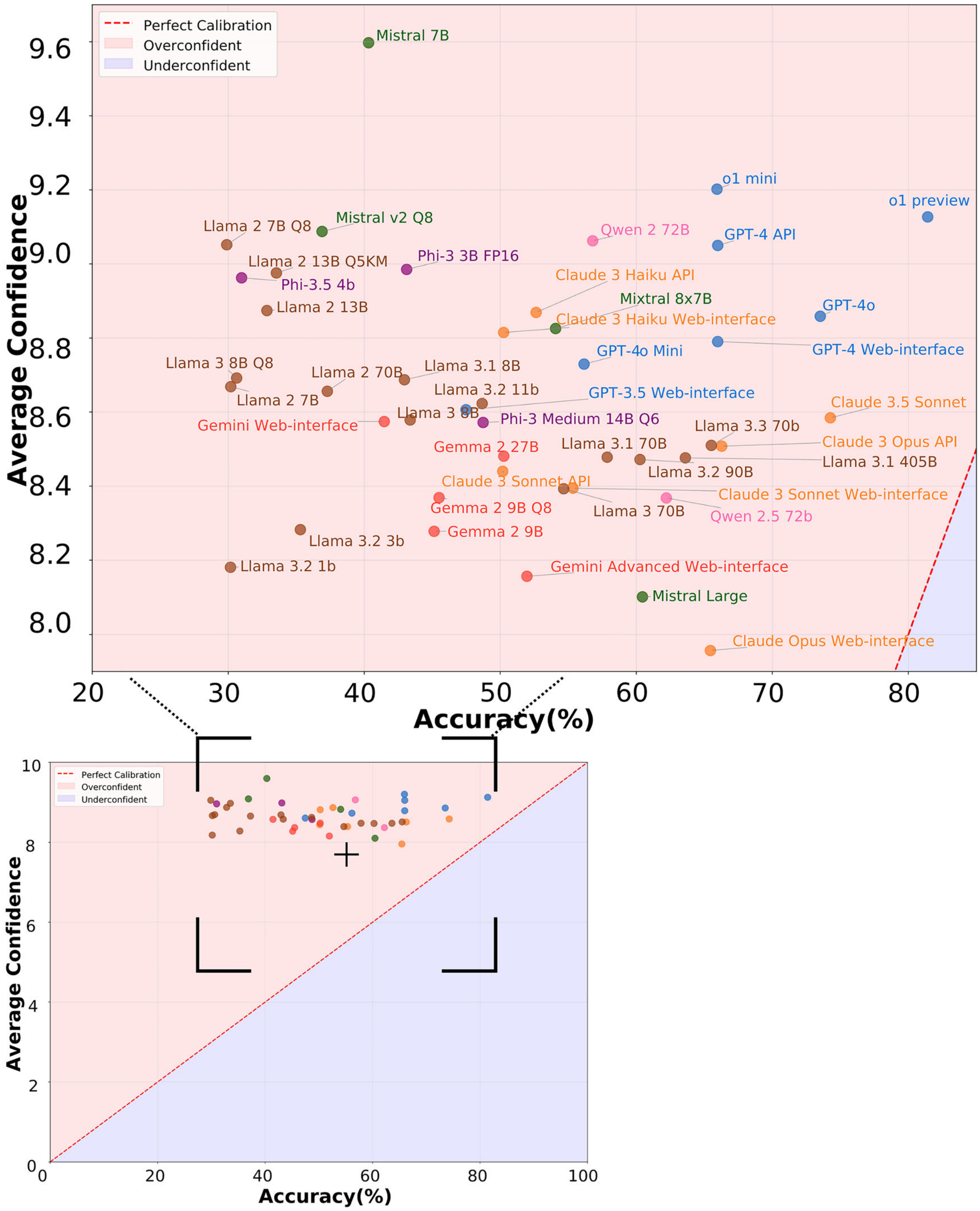

Fig. 2 | Average response accuracy versus mean self-reported confidence across large language models. Scatterplot of mean accuracy and mean confidence scores (0—10 scale) for models with more than 150 valid responses, with each point representing a single model. The dashed line denotes perfect calibration, where mean confidence equals mean accuracy. Models above this line are overconfident, whereas models below are under-confident. A subset of closely clustered models is magnified to improve readability.

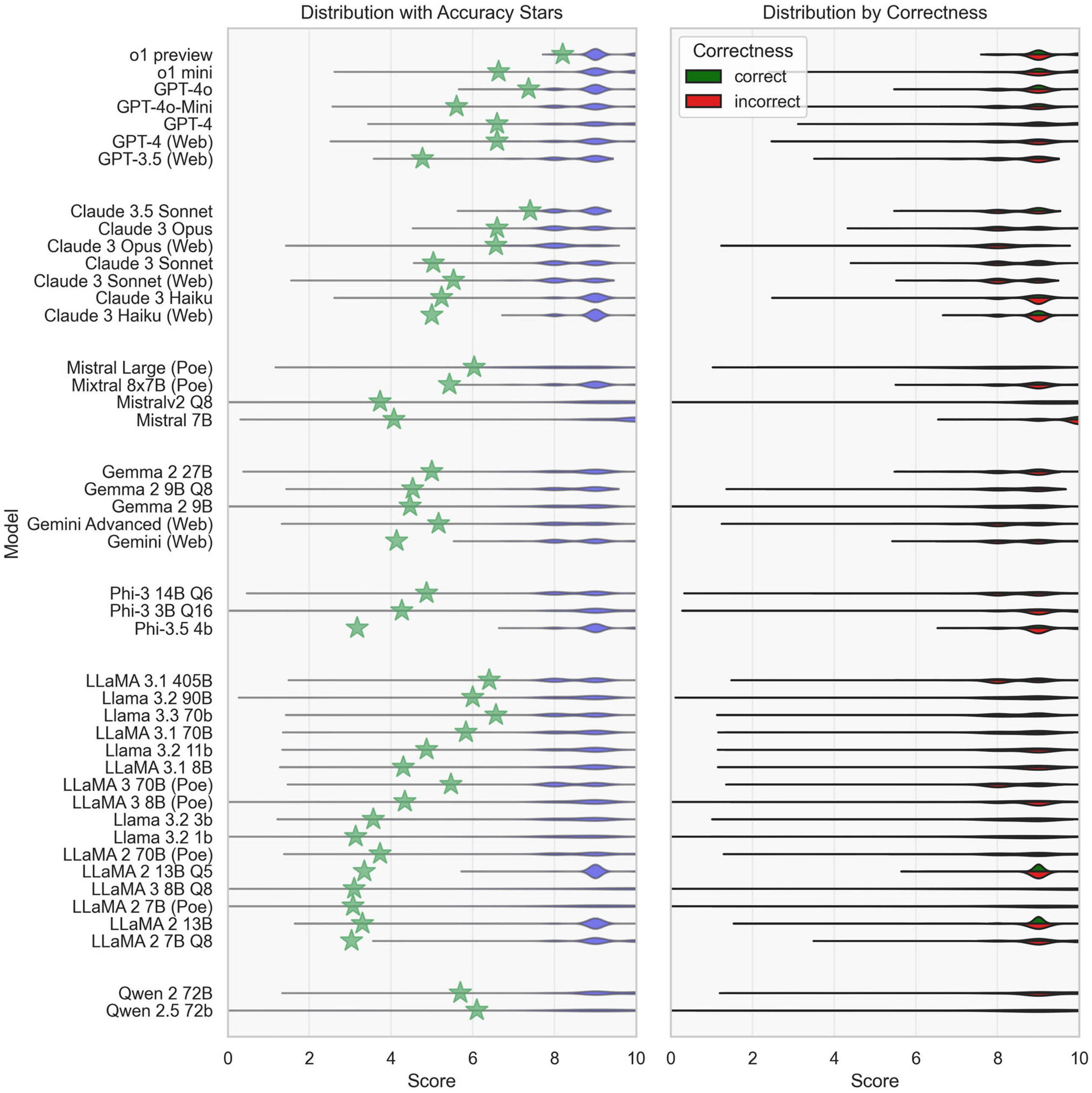

Fig. 3 | Distributions of self-reported confidence by model and correctness of response. Left panel: overall distributions of self-reported confidence scores (0−10 scale) for each model, with star markers indicating mean response accuracy for that model. Right panel: distributions of self-reported confidence scores stratified by response accuracy, showing separate curves for correct and incorrect answers for each model.

systematic overconfidence, with average confidence consistently exceeding average accuracy (Fig. 2).

We observed a substantial overlap in confidence distributions between correct and incorrect responses, indicating limited discriminative capacity (Fig. 3). Models expressed high certainty regardless of whether their answers were right or wrong—a critical safety issue in clinical settings. We quantified this observation with discrimination metrics. Even the best-performing model (o1 mini) achieved an Area Under the Receiver Operating Characteristic (AUROC) of only 0.626, well below the 0.7 threshold typically considered meaningful for clinical applications (Table 1; Supplementary Fig. S3). This pattern was consistent across all model families.

Calibration analyses corroborated these results. Only 5 of 48 models demonstrated better-than-random calibration. The best Brier scores were observed for o1-preview (0.157), followed by Claude-3.5-Sonnet (0.202) and GPT-4o (0.206) (Supplementary Fig. S4). Brier score quantifies the mean squared gap between predicted confidence and the observed outcome; lower values indicate better calibration. Calibration curves (Fig. 4, Supplementary Fig. S5) likewise confirm a consistent pattern of overconfidence. Expected Calibration Error (ECE), defined as the weighted average absolute difference between predicted confidence and empirical accuracy across confidence bins, shows the same trend, with even the best models (o1-preview: 0.100; Claude-3.5-Sonnet: 0.122) deviating meaningfully from perfect calibration (Supplementary Fig. S6).

Most alarming, we found that models maintained high confidence even as their accuracy significantly decreased on the most challenging questions for humans (Fig. 5). Even the best-calibrated models (Fig. 5a–c) showed similar overconfidence on difficult questions as poorly calibrated models (Fig. 5d–f). We also investigated whether question length affected confidence assessments, finding that confidence scores remained stable regardless of text complexity and had no meaningful relationship with

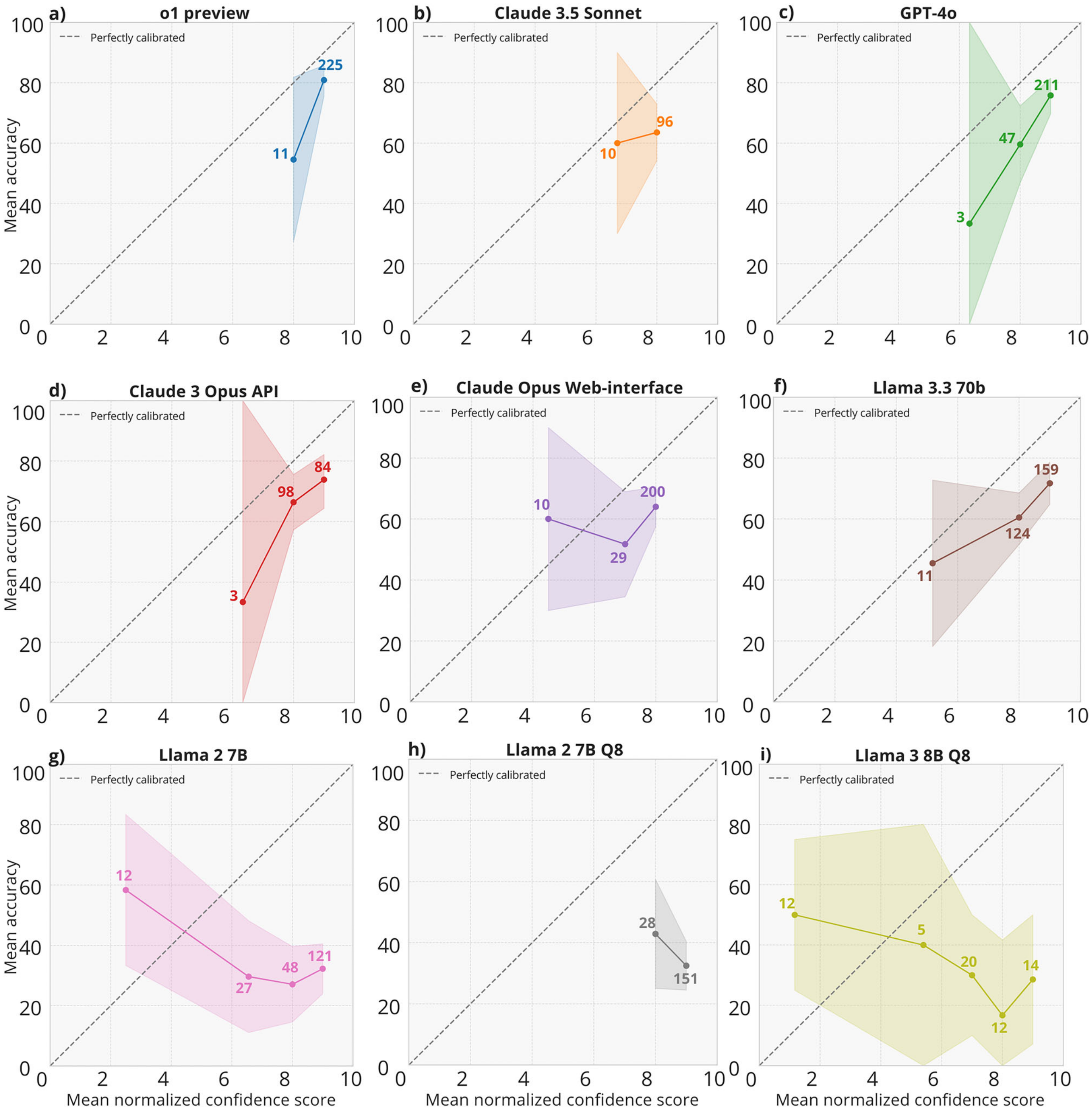

Fig. 4 | Calibration of self-reported confidence across best- and worst-calibrated large language models. a-i show calibration curves for each LLM, plotting observed model accuracy (*y*-axis, %) against mean normalized self-reported confidence (*x*-axis) within 15 quantile-based bins of the original 0—10 confidence scale. The dashed diagonal line indicates perfect calibration. Points represent confidence bins, with adjacent numbers indicating the number of questions contributing to each bin; bins containing fewer than three responses are not displayed. Panels are ordered by Brier score from best- to worst-calibrated models, with the six lowest Brier-score (best calibrated) models in **a**–**f** and the three highest Brier-score (worst calibrated) models in **g**–**i**.

actual performance (Supplementary Fig. S7). This suggests models lack the awareness to recognize that longer, potentially more complex questions could reduce their response certainty.

Looking at differences between model families, we observed generational improvements in self-assessed confidence. Newer versions consistently outperformed their predecessors. For example, o1 showed better calibration than GPT-4o, which in turn outperformed GPT-4 (Table 1). Commercial models generally demonstrated superior uncertainty estimation compared to equivalent open-source alternatives, though this pattern had notable exceptions. We also found that quantization, while enabling deployment on less powerful hardware, typically degraded calibration (as seen when comparing Llama 3 8B with its quantized counterpart). Additional analysis of middle-performing models further confirmed these trends (Supplementary Fig. S5).

Our findings confirm previous research describing the limitations of LLM self-reported confidence and provides three additional contributions[11,15]. First, we present the most comprehensive cross-architectural evaluation to date, testing 48 LLMs—from 7 B to 175 B parameters—across commercial, open-source, and quantized deployments. Second, by using gastroenterology board-style questions, we deliver key domain-specific insights. Third, we show quantitatively that all models suffer a common metacognitive deficiency, in which even the best-calibrated LLMs remain systematically overconfident, regardless of question difficulty. This pervasive overconfidence transcends architecture,

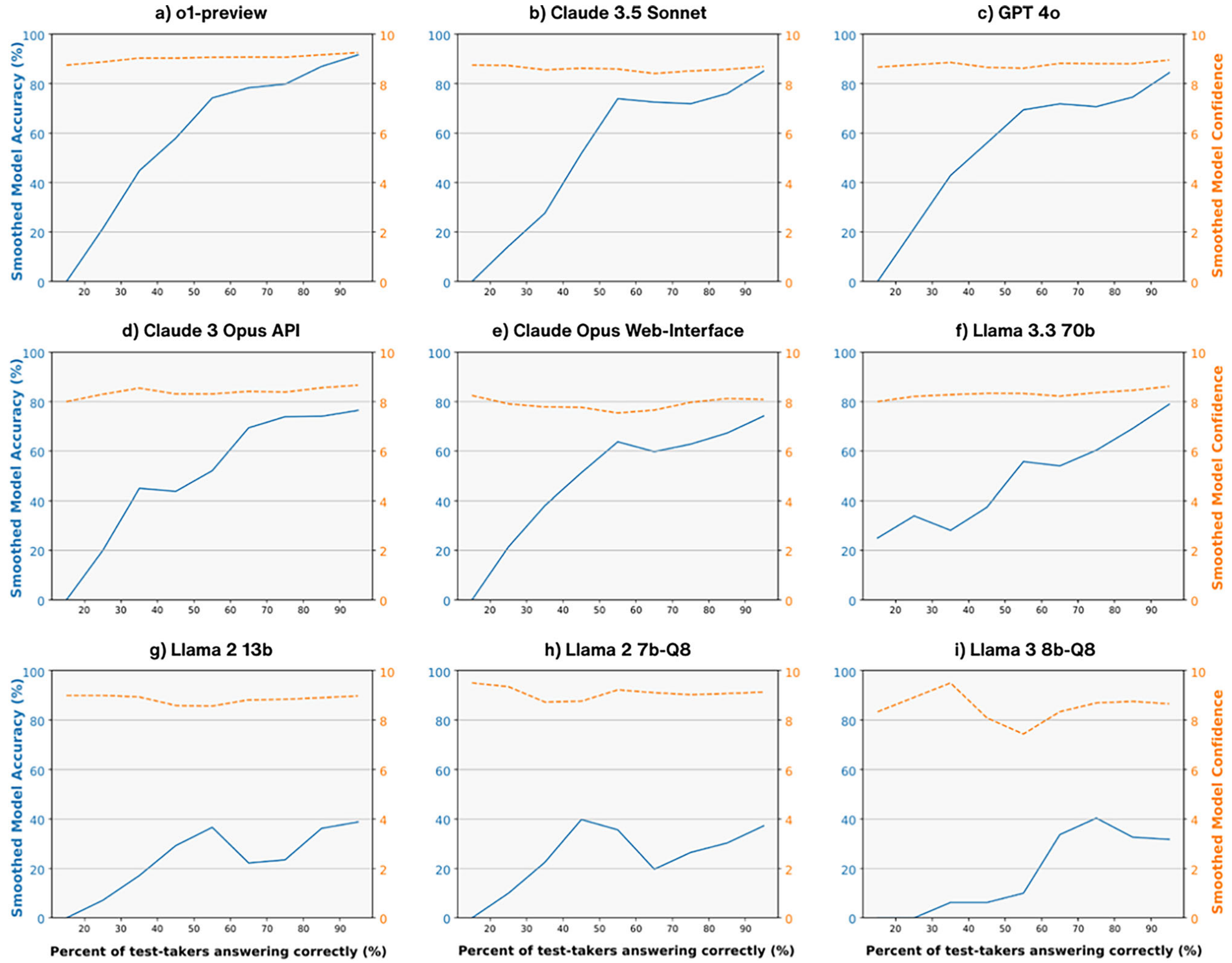

Fig. 5 | Accuracy and self-reported confidence across question difficulty for nine large language models. **a**–**i** show smoothed model accuracy (blue solid line, left *y*-axis, %) and smoothed self-reported confidence (orange dashed line, right *y*-axis, 0–10 scale) as a function of question difficulty. Question difficulty is defined as the percentage of human test-takers answering correctly (lower values indicate more difficult questions), and questions were grouped into 5-percentage-point difficulty bins on the *x*-axis. Panels are ordered by Brier score from best- to worst-calibrated models, with the six lowest Brier-score (best calibrated) models in **a**–**f** and the three highest Brier-score (worst calibrated) models in **g**–**i**.

scale, and deployment environment, pointing to a fundamental limitation of current neural language models.

While we observed that newer model generations (for example, o1 and Claude 3.5) achieve modestly better calibration metrics[11], this improved calibration tracks closely with higher overall accuracy (Table 1). These gains appear to reflect the decreased chance of inaccurate responses, rather than an improved ability to communicate uncertainty[9]. Supporting this conclusion, we observed high, unvarying self-reported confidence scores, irrespective of question difficulty (Fig. 5), model generation (Fig. 2), or correctness (Fig. 3). This suggests that confidence outputs are simply statistically generated text outputs, rather than true reflections of model uncertainty. In other words, the models are generating the most probable text for a confidence estimate, rather than reflecting on their own knowledge boundaries. Future improvements in model training and calibration may improve uncertainty self-awareness, but it is unclear if LLMs are structurally capable of this functionality.

Several limitations temper our conclusions. While our use of multiple-choice gastroenterology board exam style questions offers a clear, objective benchmark, this approach may not generalize to open-ended or multi-turn clinical reasoning or to other medical specialties. Our standardized prompt engineering approach, which was designed to maximize accuracy, could itself have biased model responses toward overconfident “expert” language. Also, we did not examine potential biases in self-reported confidence across demographic variables such as age, gender, or race, which may influence outcomes and have important real-world implications. Finally, while the ACG self-assessment questions are proprietary and only accessible to paying subscribers or via direct request, we cannot rule out that the LLMs we tested may have been trained on this data.

Despite these limitations, our findings highlight critical AI safety concerns. Across architectures, scales, and deployment settings, LLMs consistently overestimated certainty and failed to differentiate between easy and difficult questions. In high-stakes clinical settings, such miscalibration can undermine safe human-AI collaboration. While newer LLM generations demonstrate improvements in response accuracy, they continue to struggle with conveying uncertainty, reflecting a broader technological limitation in language-based metacognition. Until this gap is addressed, self-reported confidence should not be relied upon as a standalone LLM safety signal.

This comprehensive evaluation of 48 LLMs reveals that poor self-reported confidence calibration is an intrinsic limitation across all tested language models. The observed high confidence scores, regardless of accuracy, question difficulty, or model architecture, suggest verbalized confidence reflects statistically generated text patterns rather than true self-assessment. These findings, combined with similar observations in general medical tasks, indicate that self-reported confidence should not be relied upon as a standalone indicator of response reliability in gastroenterology and likely other clinical domains. Human oversight of LLM outputs remains essential. Future research should determine whether these calibration deficits extend to other medical specialties, open-ended clinical reasoning, and models specifically trained for uncertainty quantification.

Methods

Reference dataset

The 2022 American College of Gastroenterology (ACG) self-assessment consists of 300 questions, of which 138 contain images. These questions were developed by a committee of gastroenterologists to reflect the knowledge, skills, and attitudes required for excellent patient care, covering a broad range of topics, including liver, colon, esophagus, pancreaticobiliary, and endoscopy. The questions were designed to assess higher-order thinking skills and were primarily case based. They were validated through statistical analysis of test-takers' performance, with an average correctness rate of 74.52% ± 19.49% on the 2022 assessment, indicating a moderate level of difficulty. Only the text portions of the questions and answers were used in this study. Questions were categorized by length (token count), difficulty (percentage of correct answers by test-takers), and patient care phase (treatment, diagnosis, or investigation). Additional details are provided in the Supplementary Section 1.

Response generation and confidence score elicitation

For response generation and confidence score elicitation, we built upon our established methodology[12], using 60 questions from the 2023 self-assessment exam and GPT-3.5 to select the model settings (temperature, maximum input, and output token count), prompt structure, and output format of all models. The configuration that maximized response accuracy was a temperature of 1, maximum token count of input token count + 512 output tokens, structured output approach, and prompt (Fig. 1). Among the various prompt engineering techniques evaluated, the following were identified as having a positive impact on the outcomes: expert mimicry, contextual embedding, Answer and Justify, Chain of Thought, confidence scoring, and direct questioning (explained in Supplementary Section 1). The OpenAI Web interface, OpenAI API, Claude Web interface, Claude API, Gemini Web interface, Poe Web interface, Firework API, and locally hosted hardware configurations such as RTX4090Ti and H100 systems were used for response generation and confidence score elicitation.

Output parsing

To efficiently extract response and confidence data from the LLM outputs, we developed a structured output pipeline using GPT-4o (Fig. 1). Our hybrid methodology combined regex-based rules to reduce the number of input tokens and LLM-based extraction to effectively parse the key portions of the LLM outputs. The pipeline identified sentences containing "confid" for further LLM-based parsing to either extract the certainty score (0–10) or define the score as "not_mentioned". Sentences classified as "not_mentioned" in the first pass were passed through the LLM-based parsing step a second time to maximize the extraction performance. The complete output parsing methodology is described in Supplementary Section 2. To validate the output parsing pipeline, we compared it against manually extracted confidence scores from five randomly selected questions per model, achieving 98.9% accuracy (Supplementary Fig. S1).

Because some models did not reliably generate confidence scores, we excluded models that were missing confidence scores for more than 50% of questions (Medicine-Chat Q8, OpenBioLLM-7B Q8, Qwen Qwq-32b, and GPT-3.5 Turbo). Supplementary Fig. S2 describes the distribution of missing confidence scores, with 30 models having missing confidence scores. Supplementary Fig. S8 illustrates a stratified analysis of response accuracy by confidence score missingness for models with missing scores for more than one-third of the questions.

Statistical analysis

We evaluated each model's performance from two perspectives: discrimination, the ability to distinguish between correct and incorrect responses, and calibration, the alignment between predicted confidence and actual accuracy.

Discrimination was quantified using AUROC. Specifically, we designated each response as 1 (positive) if it was labeled "correct" and 0 (negative) otherwise. The confidence scores of the model ranged from 0 to 10 and served as the continuous predictor variable. We employed the roc_auc_score function from sklearn.metrics to calculate the AUROC. In practical terms, AUROC measures how well confidence scores can separate correct from incorrect answers, with 0.5 indicating random performance and 1.0 indicating perfect discrimination. Conceptually, this involves varying the decision threshold over all possible confidence values, thereby classifying the responses as positive or negative at each threshold.

Calibration was evaluated using calibration plots, Brier score, and ECE. Calibration plots were generated by dynamic based on data quantiles, creating 15 bins across the range of confidence scores, and plotting the mean predicted confidence against the observed accuracy in each bin. Bins containing fewer than three predictions were excluded to ensure the reliability of the results. Bootstrap resampling ($n = 1000$ iterations per bin) was used to derive 95% confidence intervals for each calibration point.

The combination of these metrics provided comprehensive assessment of model uncertainty estimation. The ECE complements the Brier score by directly quantifying the aggregate discrepancy between predicted probabilities and observed outcomes across bins, whereas the Brier score measures the mean squared error between predictions and true labels. As a result, the Brier score reflects both calibration (how closely predicted probabilities match observed frequencies) and refinement (the sharpness of predictions), whereas ECE focuses more directly on calibration quality. Calculating both metrics provides a more comprehensive evaluation of model performance, capturing not only how well models are calibrated, but also the overall predictive accuracy of their probability estimates.

Our development and analysis were performed using Python 3.10. LLM answers were generated and extracted using the OpenAI Python library, Ollama application (v0.4), LM studio, and Langchain (v0.2 and v0.3). Statistical analyses were conducted using SciPy (v1.13.1) and Scikit-learn (v1.5.1), with data manipulation and visualization implemented through Pandas (v2.2.2), Matplotlib (v3.9.2), and Seaborn (v0.13.2). Additional methodological details and code are available in our repository (see Code availability).

Ethical approval

This study did not require ethical approval, as it did not involve human subjects or human data. We ensured data protection by confirming that the utilized LLM services did not retain or use our queries for model training purposes.

Data availability

The data supporting this study's findings were obtained from the American College of Gastroenterology (ACG) under license agreement. While these data are not publicly available owing to licensing restrictions, they may be obtained from the authors with the ACG's permission upon reasonable request. ACG self-assessment questions and answers are accessible to members through https://education.gi.org/.

Code availability

The underlying code for this study is available at https://github.com/narimannr2x/confidence_scoring.

Received: 22 May 2025; Accepted: 7 January 2026;
Published online: 05 February 2026

References

1. Liévin, V., Hother, C. E., Motzfeldt, A. G. & Winther, O. Can large language models reason about medical questions? *Patterns* **5**, 100943 (2024).
2. Singhal, K. et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **31**, 943–950 (2025).
3. Haltaufderheide, J. & Ranisch, R. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *Npj Digit. Med.* **7**, 183 (2024).
4. McKenna, N. et al. Sources of hallucination by large language models on inference tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023* 2758–2774 (Association for Computational Linguistics, 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.182.
5. Xiong, M. et al. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *International Conference on Learning Representations (ICLR 2024)* (2024). https://doi.org/10.48550/arXiv.2306.13063.
6. Duan, J. et al. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* 5050–5063 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.acl-long.276
7. Wu, J., Yu, Y. & Zhou, H.-Y. Uncertainty estimation of large language models in medical question answering. Preprint at https://doi.org/10.48550/arXiv.2407.08662 (2024).
8. Tian, K. et al. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. in *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* 5433–5442 (Association for Computational Linguistics, 2023). https://doi.org/10.18653/v1/2023.emnlp-main.330.
9. Yang, D., Tsai, Y.-H. H. & Yamada, M. On Verbalized Confidence Scores for LLMs. (2024). https://doi.org/10.48550/arXiv.2412.14737.
10. Ni, S., Bi, K., Yu, L. & Guo, J. Are Large Language Models More Honest in Their Probabilistic or Verbalized Confidence? In *Information Retrieval. CCIR 2024.* Lecture Notes in Computer Science, vol. 15418 (Springer, Singapore, 2025). https://doi.org/10.1007/978-981-96-1710-4_10.
11. Omar, M., Agbareia, R., Glicksberg, B. S., Nadkarni, G. N. & Klang, E. Benchmarking the confidence of large language models in clinical questions. *JMIR Med. Inform.* **13**, e66917 (2025).
12. Savage, T. et al. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *J. Am. Med. Inform. Assoc. JAMIA* **32**, 139–149 (2025).
13. Griot, M., Hemptinne, C., Vanderdonckt, J. & Yuksel, D. Large Language Models lack essential metacognition for reliable medical reasoning. *Nat. Commun.* **16**, 642 (2025).
14. Safavi-Naini, S. A. A. et al. Benchmarking proprietary and open-source language and vision-language models for gastroenterology clinical reasoning. *npj Digit Med* **8**, 797 (2025).
15. Vashurin, R. et al. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Trans. Assoc Comput. Linguist* **13**, 220–248 (2025).

Acknowledgements

This study was supported by the American Gastroenterological Association AGA-Amgen Fellowship-to-Faculty Transition Award (AGA2023-32-06) for AS. The funding source had no role in the study design, data collection, analysis, interpretation, or manuscript preparation. We thank the American College of Gastroenterology for providing their question bank, the Hugging Face team for their accessible AI infrastructure, and the Bloke account on Hugging Face for providing quantized versions of open-source LLMs. ChatGPT was used to assist with English language editing during manuscript preparation. The authors reviewed and edited all AI-assisted content and maintained full responsibility for the manuscript's content.

Author contributions

N.N.: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Writing Original Draft, Programming, Data Curation; S.A.A.S.N.: Methodology, Investigation, Validation, Review & Editing, Project Administration; A.S.: Methodology, Investigation, Supervision, Validation; TS: Investigation, Validation; G.N.: Supervision; Z.A.: Investigation; P.L.: Investigation; M.K.: Investigation.

Competing interests

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this study. N.N.: none; S.A.A.S.N.: none; T.S.: none; M.K.: none; P.L.: none; Z.A.: none; G.N.: is a founder of Renalytix, Pensieve, and Verici and provides consultancy services to AstraZeneca, Reata, Renalytix, and Pensieve. He also has equity in Renalytix, Pensieve, and Verici; A.S. is on the advisory board and has equity in Virgo Surgical Solutions.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44355-026-00053-3.

Correspondence and requests for materials should be addressed to Ali Soroush.

Reprints and permissions information is available at http://www.nature.com/reprints