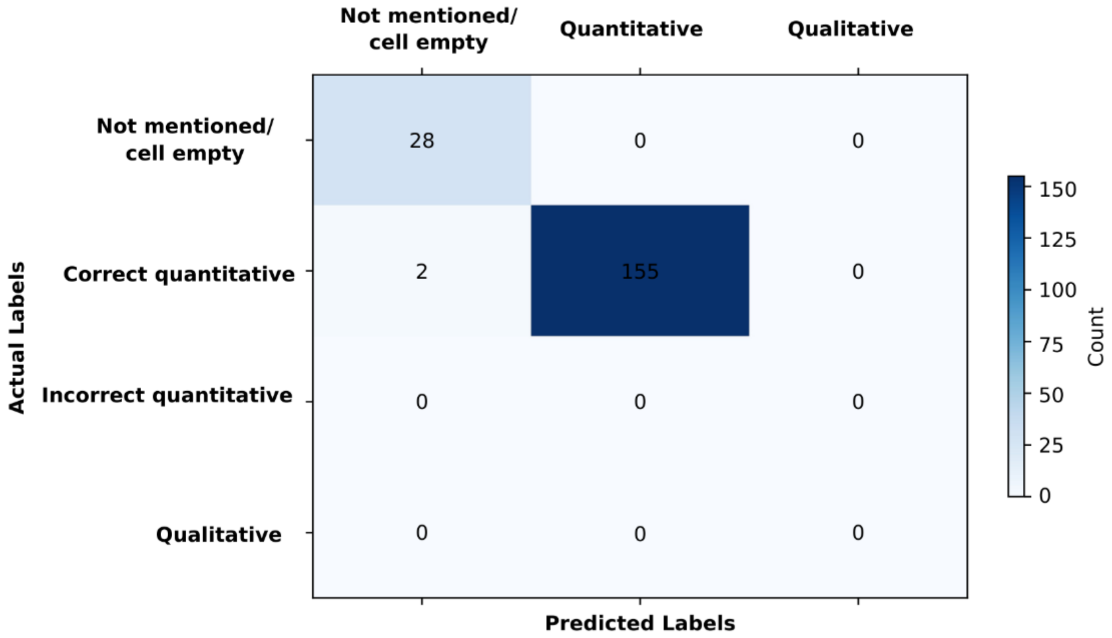
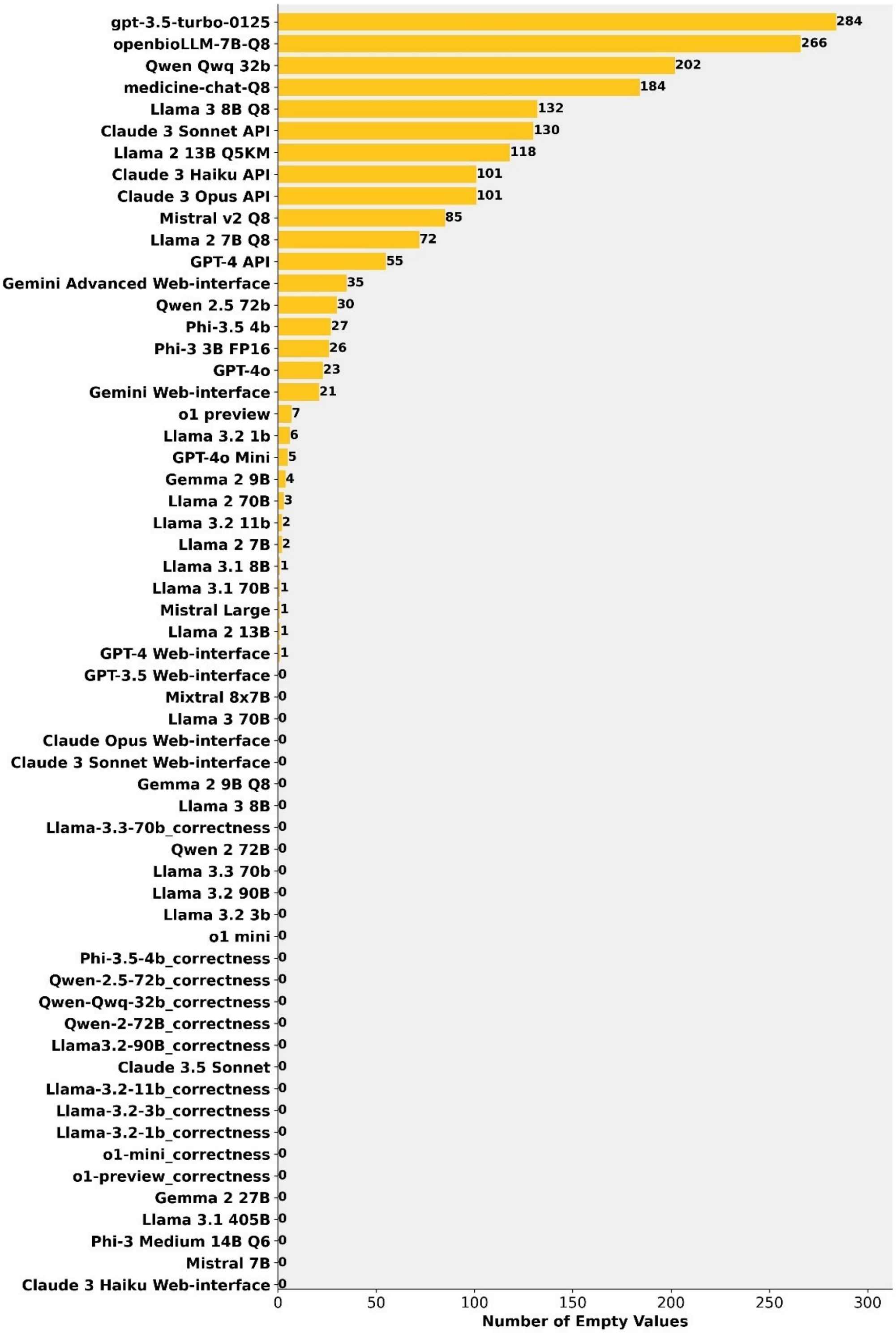
Supplementary Data

This is a supplementary file to "**Across Generations, Sizes, and Types, Large Language Models Poorly Report Self-Confidence in Gastroenterology Clinical Reasoning Tasks**" by Nariman Naderi, Seyed Amir Ahmad Safavi-Naini, Thomas Savage, Mohammad Amin Khalafi, Zahra Atf, Peter Lewis, Girish Nadkarni, Ali Soroush.

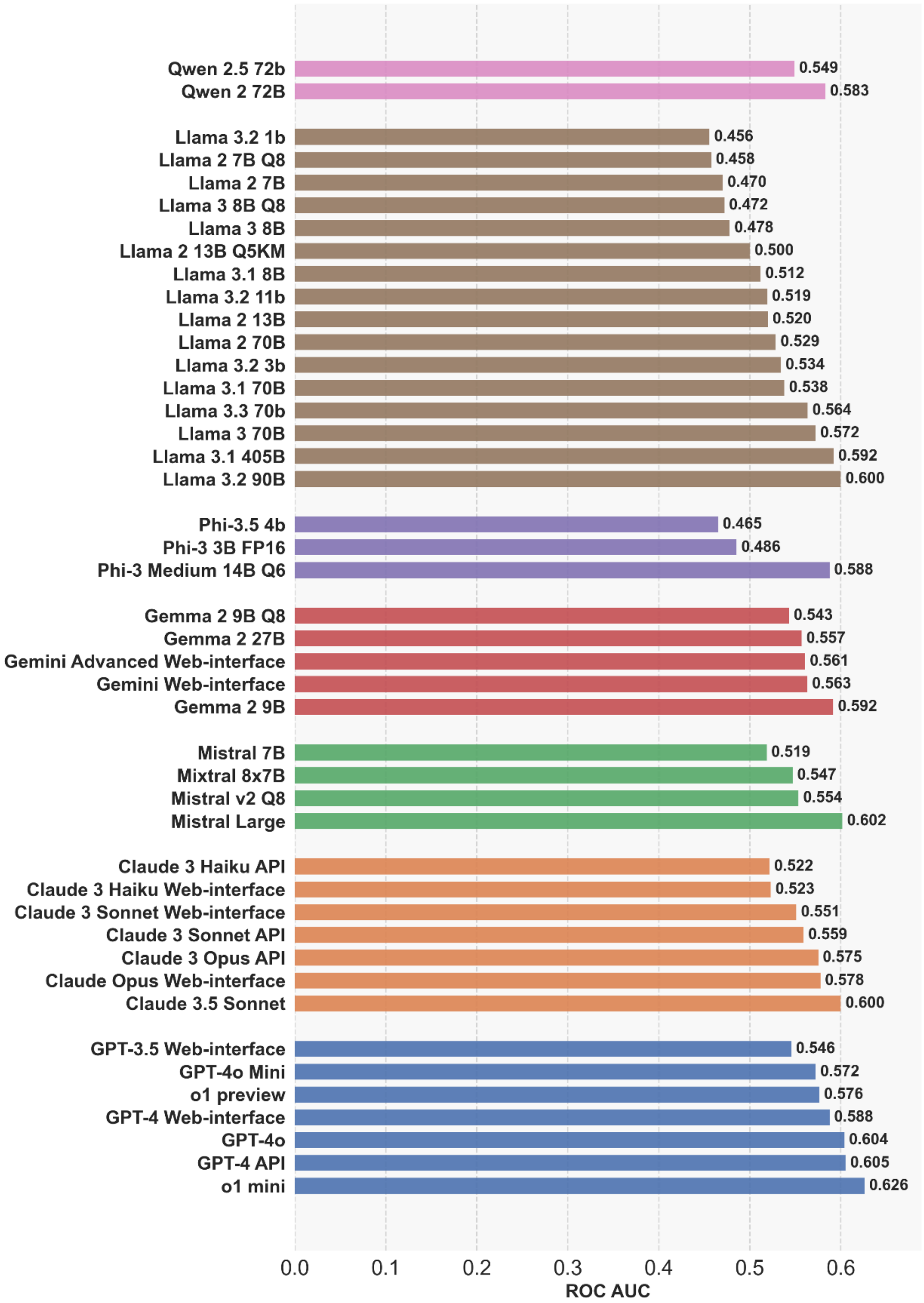
Corresponding author: Ali Soroush (Ali.Soroush@mountsinai.org).



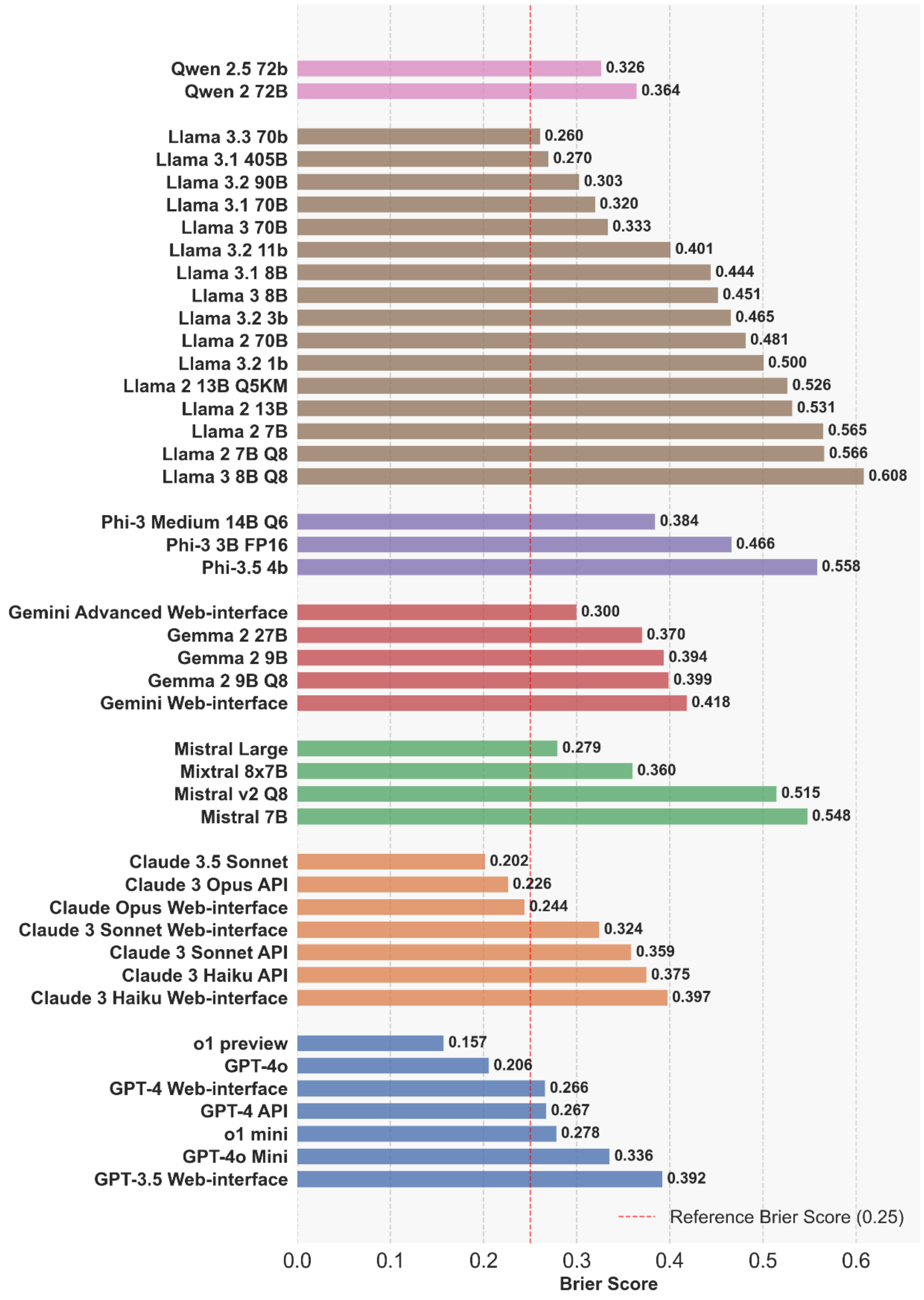
Supplementary Figure S1. Confusion matrix of the accuracy of the automatic confidence extraction pipeline. As mentioned in the text, we used an LLM extraction pipeline to extract the confidence numbers. Five questions were chosen from each model's answers for human evaluation. As stated above, the model accuracy was 98.91% (153 out of 155 questions).



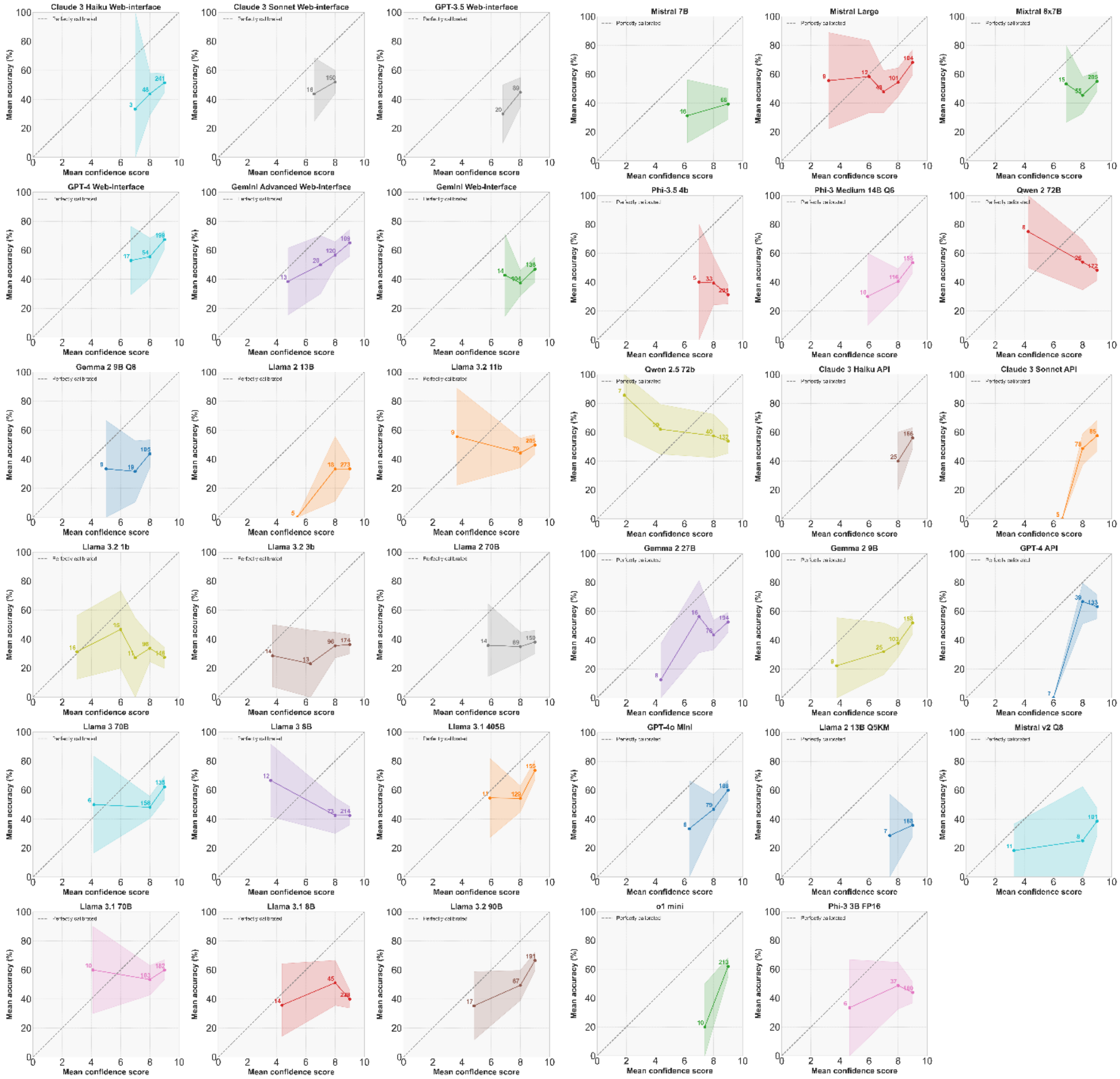
Supplementary Figure S2. Non-generated confidence elicitation for each model, sorted from highest (top) to lowest (bottom). Gpt-3.5-turbo-0125 exhibited the highest number of non-generations (n=284, 94.7%), followed by openbioLLM-7B-Q8 (n=266, 88.7%), and medicine-chat-Q8 (n=184, 61.3%).



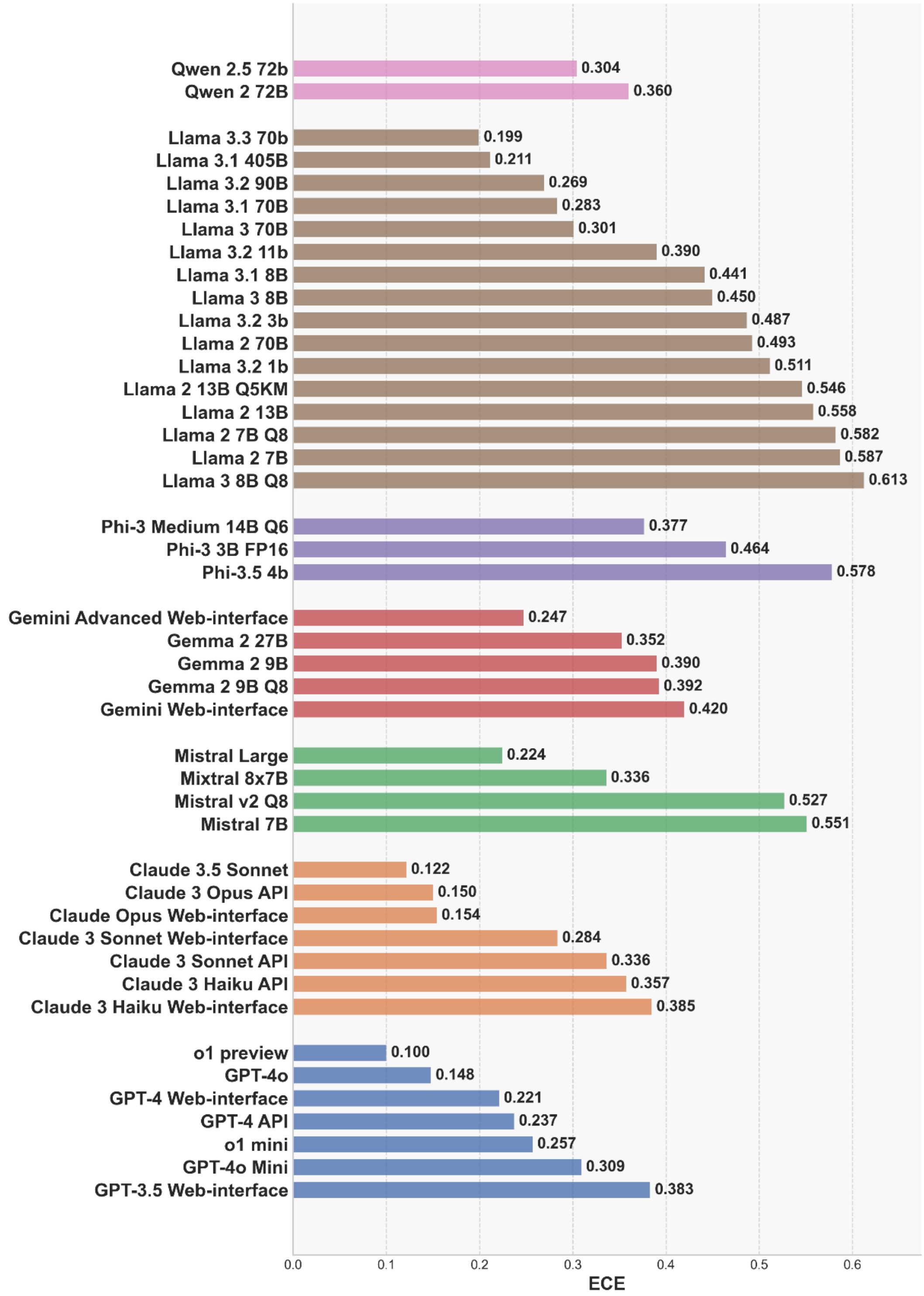
Supplementary Figure S3. AUROC Comparison of Various Model Families Based on Confidence Scores and Question Accuracy. This graph presents the AUROC for each model, reflecting their performance in assigning confidence scores to questions relative to their accuracy. The models were grouped by their respective families for easier comparison. Receiver operating characteristic (ROC) curves were generated by comparing the model-derived confidence scores with binary correctness labels, and the area under the curve was computed to evaluate model discrimination.



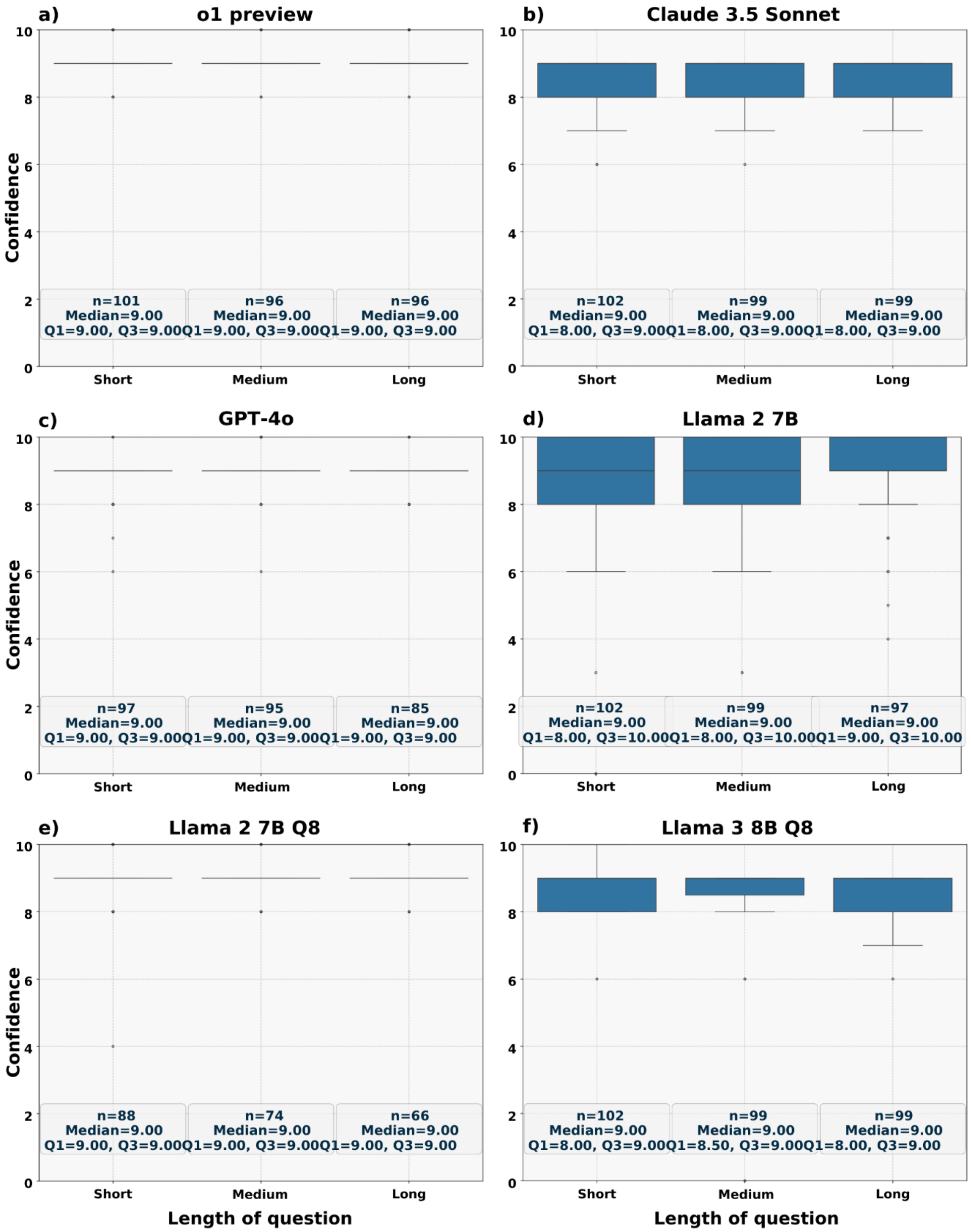
Supplementary Figure S4. Brier Scores for LLM Confidence Elicitation. The chart illustrates the comparative performance of the different language models, with lower scores indicating better calibration. The red dashed line indicates a reference Brier score of 0.25, representing the score expected from the random predictions. The models were grouped by their respective families for easier comparison.



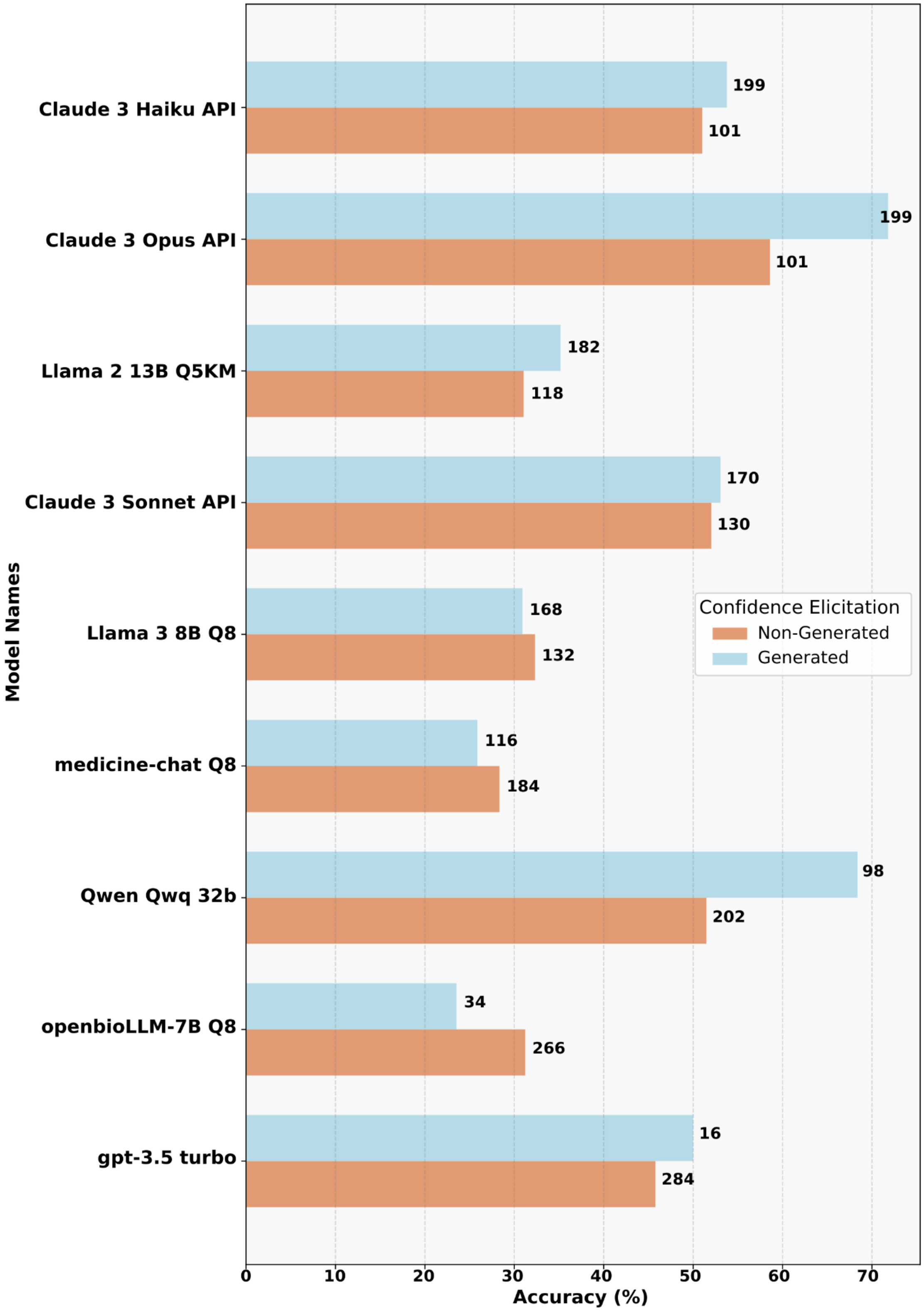
Supplementary Figure S5. Calibration Curves for Middle-Tier Models. This figure presents calibration curves for models with performance falling between the top six and bottom three models. Confidence scores were binned into intervals of 0.1 across the range of 0 to 1, with the mean normalized confidence score for each bin plotted against the corresponding observed accuracy. The dashed line represents perfect calibration.



Supplementary Figure S6. Expected Calibration Error (ECE) scores for LLMs in the context of confidence elicitation. Lower scores indicate better calibration. Although there is no universally accepted threshold, an ECE value below 0.1 is commonly regarded as acceptable. Models are grouped by their respective families to facilitate comparison.



Supplementary Figure S7. Figures (a) to (f) present box plots illustrating the confidence scores elicited by the selected models stratified by question length. Response confidence scores appear qualitatively independent of the question length. Figures (a)–(c) highlight the three models with the lowest Brier scores (highest calibration), whereas Figures (d)–(f) display the three models with the highest Brier scores (lowest calibration).



Supplementary Figure S8. Model accuracy stratified by the generation of confidence elicitation.

Supplementary Section 1: Exam and Question Context

This section provides additional details regarding the dataset used for evaluating LLM and VLM performance and the methodology employed for parsing LLM outputs, as referenced in the main text.

1.1 ACG Self-Assessment Dataset Characteristics

The primary dataset utilized in this study is the 2022 American College of Gastroenterology (ACG) Self-Assessment Test (SA). This examination is meticulously developed by an ACG committee, incorporating contributions from postgraduate course faculty members, to reflect the knowledge, skills, and attitudes essential for excellent patient care in gastroenterology.

- **Content and Scope:** The 2022 ACG-SA comprises 300 multiple-choice questions covering a broad spectrum of gastroenterology topics, including but not limited to the liver, colon, esophagus, pancreaticobiliary system, and endoscopy procedures. Of these 300 questions, 138 include associated images (e.g., endoscopy, radiology, histology).
- **Design and Cognitive Level:** The questions are primarily case-based (297 out of 300 in the 2022 set) and are designed to assess higher-order thinking skills, moving beyond simple recall to evaluate clinical reasoning and application of knowledge, aligning with Level 2 of a modified Bloom's taxonomy.
- **Validation and Difficulty:** The questions undergo validation through statistical analysis based on the performance of actual test-takers (gastroenterologists and fellows-in-training). For the 2022 assessment, the average correctness rate among human test-takers was 74.52% ± 19.49%, indicating a moderate overall difficulty level.
- **Data Usage in This Study:** For the analyses involving text-only LLMs (Experiment 1), only the textual portions of the questions and their corresponding multiple-choice answers were used. Image data was explicitly excluded for these models but utilized in VLM assessments (Experiment 2).
- **Question Categorization:** To facilitate stratified performance analysis, questions were categorized based on several characteristics:
    - **Difficulty:** Defined by the percentage of human test-takers who answered correctly. Questions were divided into four quartiles: Q1 (most difficult: 12.75%-64.92% correct), Q2 (64.93%-79.23% correct), Q3 (79.23%-89.44% correct), and Q4 (easiest: 89.45%-99.21% correct).
    - **Length:** Measured by the total token count (question stem + options) using the tiktoken library. Questions were divided into three tertiles: Short (49-179 tokens), Medium (180-262 tokens), and Long (263-588 tokens).
    - **Patient Care Phase:** Classified based on the primary focus of the question, including Diagnosis (n=123), Treatment (n=217), Investigation (n=211), Management of Complications (n=55), or Pathophysiology (n=3). Note that questions could be tagged with multiple phases.
    - **Subject:** Categorized by the ACG into specific gastroenterology topics (e.g., Liver, Colon, Esophagus, IBD, Endoscopy, etc.).

1.2 LLM Output Parsing Methodology for Confidence Score Extraction

To efficiently and accurately extract structured information, specifically the LLM's self-reported confidence score, from potentially unstructured textual outputs, we developed and implemented a dedicated pipeline leveraging GPT-4o, as illustrated conceptually in Figure 1 of the main text. This was particularly relevant for extracting the confidence rating requested by our optimized prompt ("...rate your confidence in this decision from 1 to 10...").

The pipeline employed a hybrid methodology combining regular expressions (regex) and LLM-based parsing:

1. **Regex-based Pre-filtering:** Initial processing used regex rules to identify sentences or phrases within the LLM's generated text that contained variations of the word "confidence" (e.g., "confid", "confident", "confidence is"). This step served to significantly reduce the number of tokens requiring further, more computationally intensive LLM-based analysis, thereby improving efficiency.

2. **LLM-based Extraction (First Pass):** Sentences identified by the regex filter were passed to an LLM parser (developed using GPT-4o capabilities). This parser was tasked with extracting a numerical confidence score within the specified range (0-10). If a score was successfully extracted, it was recorded. If the LLM parser could not identify a score within the targeted sentence despite the presence of "confid*", the sentence was flagged as "not_mentioned".

3. **LLM-based Extraction (Second Pass):** Sentences initially classified as "not_mentioned" underwent a second round of LLM-based parsing. This re-parsing step aimed to maximize extraction performance by attempting to capture confidence scores that might have been phrased in less direct ways or located in slightly different contexts within the sentence, which the first pass might have missed. If a score was found in the second pass, it was recorded; otherwise, the confidence for that response remained classified as "not_mentioned".

This multi-step, hybrid approach allowed for robust and efficient extraction of the self-reported confidence data from the LLM outputs for subsequent analysis. Extraction of the *selected answer option* (A, B, C, D, E) from unstructured outputs was handled separately as part of the evaluation pipeline.