**Benchmarking Hallucination Detection and Calibration in Medical LLMs: A Comparative Study of GPT-4o-mini and GPT-5-mini**

Ethan He

Providence Day School, Charlotte, NC, USA

**\* Corresponding author**  
E-mail: tiancheng.ethan.he@gmail.com (EH)  
Phone: 216-287-9026  
Address: 7137 Burlwood Rd., Charlotte, NC 28211, USA.  
**Word count:** 2364  
**Funding:** The author(s) received no specific funding for this work.  
**Conflicts of interest:** All authors declare no conflict of interest.

# Abstract

Large language models (LLMs) are increasingly applied to medical domains, yet their tendency to hallucinate—to produce incorrect outputs—raises concerns for clinical reliability [1, 2]. In this study, we evaluate hallucination behavior in GPT-4o-mini and GPT-5-mini using the MedQuAD: Medical Question–Answer Dataset of 14,984 curated medical question–answer pairs [3]. Each model was prompted to generate responses with self-reported confidence scores, which we assessed against gold-standard answers using a state-of-the-art natural language inference (NLI) classifier [4].

Analyses compared hallucination rates across biomedical sectors, examined cross-model agreement, and tested whether self-reported confidence tracked factual accuracy. We further explored hallucination severity and the relationship between model confidence and external entailment probabilities.

GPT-5-mini substantially outperformed GPT-4o-mini, reducing hallucination rates from 49.6% to 31.1% (18.5% absolute improvement). Gains were concentrated in domains such as neurology and kidney disorders, while overlap analysis showed GPT-5 corrected 28.5% of GPT-4’s failures but regressed in 10%. Calibration also improved, with fewer high-confidence hallucinations, though overconfidence persisted.

These findings demonstrate measurable but incomplete progress from GPT-4 to GPT-5. By integrating sector, severity, and calibration, our evaluation framework provides a reproducible approach for benchmarking medical LLMs and highlights the continued need for safeguards in clinical deployment.

**Keywords:** Large language models; hallucination; calibration; medical question answering; clinical safety

# 1 Introduction

Large language models (LLMs) have rapidly transitioned from research prototypes to tools with emerging applications in healthcare. Their ability to generate fluent text, synthesize literature, and answer patient-facing queries suggests real potential for supporting diagnostics, clinical communication, and medical writing [5]. Yet this promise is undermined by a persistent limitation: LLMs are prone to hallucinations—outputs that sound confident but are factually incorrect. In clinical contexts, where accuracy is essential, such errors pose unacceptable risks to patient safety and trust in AI-assisted care [1, 2].

Early evaluations illustrate this duality. Although ChatGPT has demonstrated passing-level performance on the United States Medical Licensing Examination, detailed analysis shows frequent factual errors, highlighting limitations for real-world clinical use [6]. Other assessments show that while LLMs often tend to generate detailed responses, their accuracy is inconsistent across medical domains [2]. These findings underscore the need for evaluation methods that go beyond average accuracy to identify when, where, and how hallucinations occur.

A commonly proposed strategy for benchmarking hallucinations involves using self-reported confidence scores. Although these scores certainly provide a signal of reliability, they cannot yet be relied on as a sole measure of factual accuracy. Models often remain miscalibrated, with high-confidence errors occurring frequently [7, 8]. This overconfidence is especially problematic in clinical contexts, where misleading certainty can undermine trust and increase risk. As such, key questions remain: Are newer models better calibrated? Which medical domains are most prone to hallucination? And have newer generations actually achieved meaningful accuracy gains?

To address these questions, we evaluate hallucination behavior in GPT-4o-mini and GPT-5-mini using the MedQuAD dataset of nearly 15,000 curated medical question–answer pairs from authoritative sources [3]. Responses were assessed with an NLI classifier, which labeled each output as *entailment*, *contradiction*, or *neutral*. This approach enables fine-grained distinctions between correct outputs and hallucinations, while also allowing calibration analysis through comparison with models’ self-reported confidence.

This study has three main objectives: (1) quantify hallucination rates across biomedical sectors; (2) compare GPT-4o-mini and GPT-5-mini to assess whether newer models reduce error frequency; and (3) test how self-reported confidence aligns with correctness. Together, these perspectives provide a comprehensive view of medical LLM reliability, highlighting both measurable progress and the persistent risks that must be managed before clinical deployment.

# 2 Methods

## 2.1 Dataset

This study used the MedQuAD (Medical Question–Answer) dataset [3], which contains 14,984 unique curated medical question–answer pairs across diverse biomedical domains. Each entry includes a patient-oriented question, an authoritative expert-provided answer, and metadata specifying the biomedical focus area and source (e.g., NINDS, NIDDK, GARD). The dataset was selected for its breadth of coverage and reliance on trusted biomedical resources, making it well-suited for benchmarking factual reliability in large language models (LLMs). Since the data are publicly available and contain no personally identifiable information, no ethical review was required. Basic preprocessing was applied before analysis. Duplicate questions—identified as exact word-for-word matches—were removed to avoid inflating model performance on repeated items. Metadata fields were preserved to enable later stratified analyses.

## 2.2 Models and Prompting

Two OpenAI models were evaluated: GPT-4o-mini and GPT-5-mini. These compact variants of OpenAI’s flagship series are optimized for efficiency while maintaining strong performance. They were chosen for analysis because, at the time of this study, they represented the most recent models freely accessible to users, making them both widely available and practically relevant for real-world applications. Each MedQuAD question was presented to both models using a standardized medical-expert prompt designed to elicit concise, evidence-based responses. To ensure concise responses, outputs were capped at 100 words. In addition to their textual response, models were instructed to provide a self-reported confidence score between 0.00 and 1.00 (two decimal places).

## 2.3 Hallucination Detection

Factual consistency was evaluated using the MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli model [4], a natural language inference (NLI) classifier based on the robust DeBERTa-v3 architecture [9]. This model was chosen for its state-of-the-art performance in natural language inference, having achieved top scores on multiple NLI benchmarks [4]. For each question–answer pair, the classifier compared the LLM response to the expert reference and assigned one of three labels (Table [1](#tab:nli_labels)).

**Table 1:** NLI classification labels used for hallucination detection.

|  |  |
| --- | --- |
| **Label** | **Definition** |
| Entailment | LLM response is consistent with and logically follows from the expert answer |
| Neutral | LLM response is neither supported nor contradicted by the expert answer (treated as mild hallucination). |
| Contradiction | LLM response directly conflicts with the expert answer (treated as severe hallucination). |

This schema allowed us to distinguish between correct answers, mild hallucinations, and severe hallucinations. To reduce false mismatches due to terminological variation, responses were expanded with medical synonyms before classification. In addition to categorical labels, the classifier outputs confidence probabilities for each class, which were later compared against LLM self-reported confidence to evaluate calibration and external validity.

## 2.4 Analysis Plan

Analyses were organized into four focus areas (Table [2](#tab:analysis_plan)).

**Table 2:** Summary of analysis objectives and methods.

|  |  |
| --- | --- |
| **Focus Area** | **Methods** |
| Domain-level performance | Compared hallucination rates by biomedical sector with 95% CIs. Highlighted gains in domains where GPT-4o-mini performed worst, and contrasted with the 15 most represented sectors. |
| Cross-model evaluation | Used a agreement matrix to compare outcomes between models, calculated improvement vs. regression rates, and examined shifts in NLI error types (entailment, contradiction, neutral). |
| Confidence and calibration | Assessed calibration with confidence–accuracy curves, bin-level error gaps, and extreme high- and low-confidence cases. |
| Confidence alignment | Correlated self-reported confidence with NLI entailment probabilities to test consistency between internal and external measures of certainty. |

These four areas were chosen to align with key gaps identified in prior work. Domain-level analysis tests whether improvements extend beyond aggregate accuracy into specific medical sectors, where earlier models often failed unevenly [10]. Cross-model evaluation directly addresses whether newer generations correct prior errors. Confidence and calibration analysis probes whether self-reported certainty can be a reliable proxy for factual accuracy, a critical issue for safe deployment. Finally, comparing model confidence with external NLI probabilities provides an independent check on reliability, ensuring that observed patterns are not artifacts of self-reporting alone.

## 2.5 Software and Reproducibility

All preprocessing, model interactions, and analyses were implemented in Python version 3.12.4. Statistical analyses and visualization used standard scientific libraries including NumPy, pandas, SciPy, and Matplotlib. GPU acceleration was employed to improve NLI inference speed. All scripts and analysis notebooks have been archived and are available in a public GitHub repository to ensure reproducibility.

# 3 Results

## 3.1 Cross-model Evaluation

We first compared GPT-4o-mini and GPT-5-mini on matched MedQuAD items to test whether the newer model corrected prior errors or introduced regressions.

As shown in Figure [1a](#fig:crossmodel), both models answered 40.4% of items correctly and hallucinated 21.1%, suggesting that a substantial subset of questions remains intrinsically challenging, likely due to ambiguity or sparse representation in training data. The off-diagonal cells reveal a notable asymmetry: GPT-5-mini corrected 28.5% of GPT-4o-mini’s hallucinations but regressed on only 10.0%. This imbalance demonstrates a net improvement, with GPT-5-mini resolving far more errors than it introduced. Figure [1b](#fig:crossmodel) summarizes this trend, showing a net gain of 18.5 percentage points.

|  |  |
| --- | --- |
| **Figure 1a:** Agreement matrix showing joint outcomes: both correct, both hallucinated, or one model hallucinated. | **A graph showing different colored squares  AI-generated content may be incorrect.Figure 1b:** Improvement vs. regression summary (GPT-5 fixes vs. GPT-5 regressions; net +18.5 pp). |

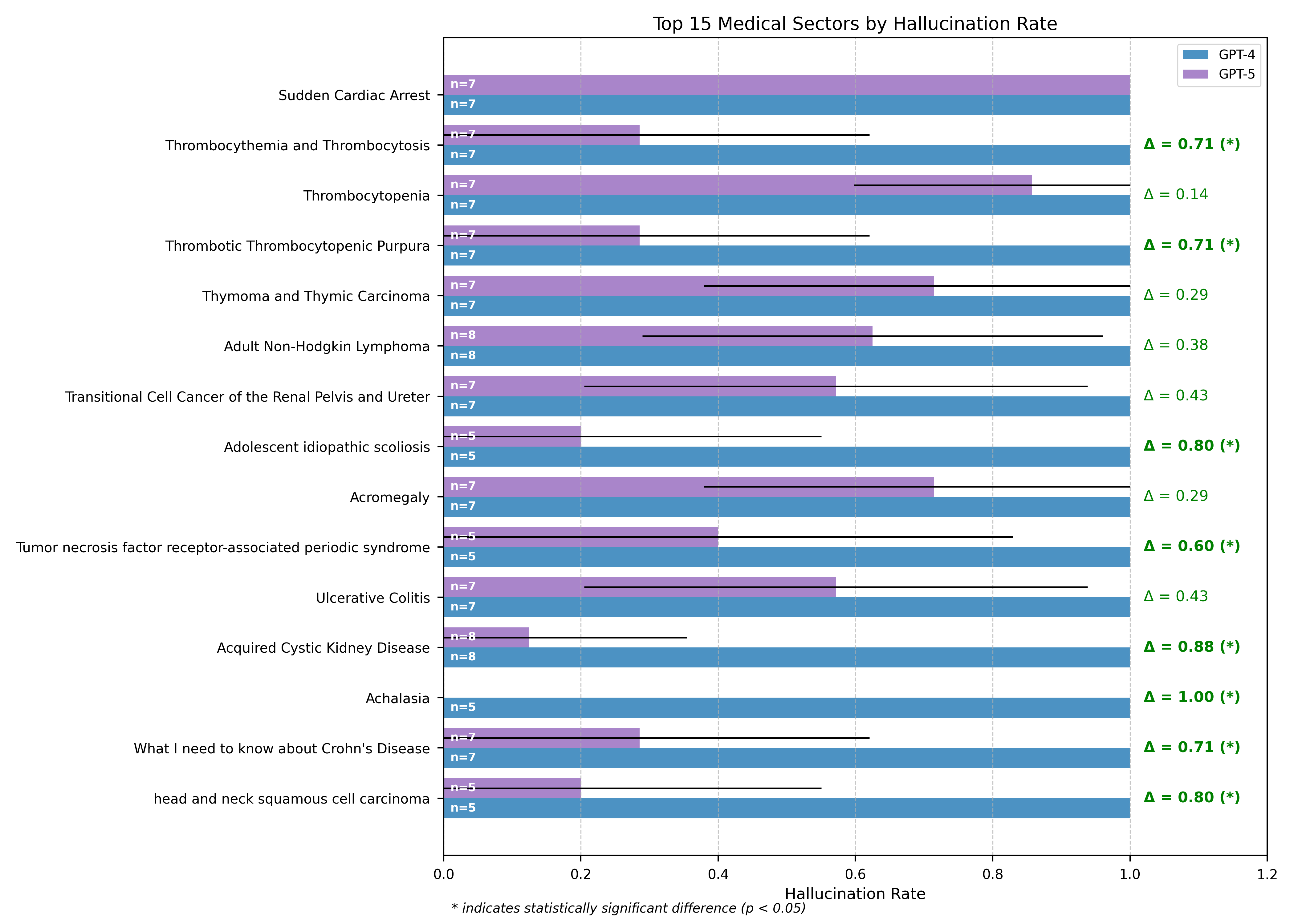
A closer look at the distribution of natural language inference categories (Figure [2](#fig:classdist)) shows that accuracy gains were concentrated in reduced contradiction errors, accompanied by increases in entailment. Neutral responses remained comparatively stable. This suggests that improvements were driven by fewer direct conflicts with reference answers rather than broad shifts across categories.

A graph of different colored bars

AI-generated content may be incorrect.**Figure 2:** Distribution by NLI classification type. GPT-5-mini increased entailment and reduced contradiction, with a small share in neutral

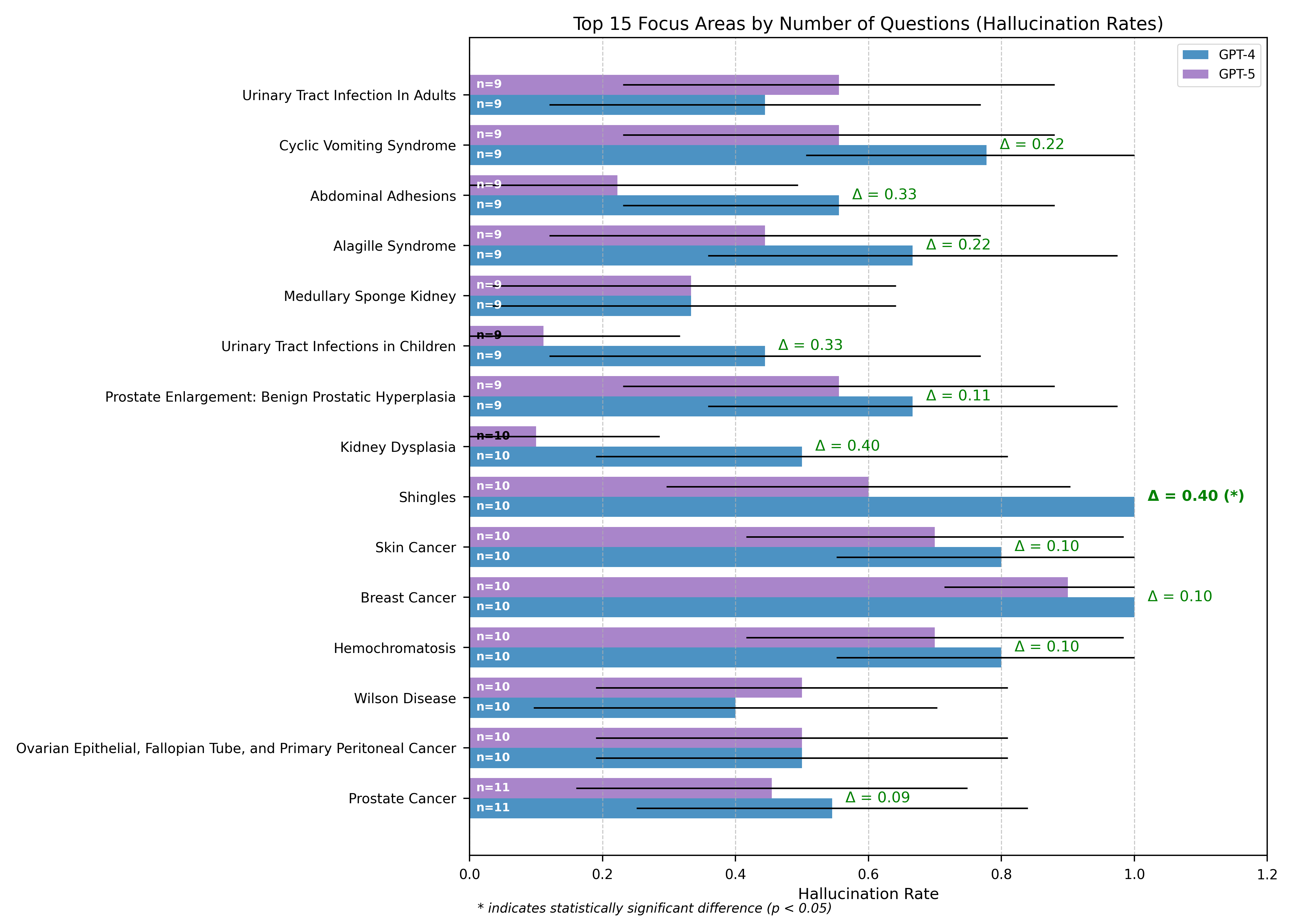
## 3.2 Domain-level Performance

Hallucination rates also varied substantially across biomedical sectors. As seen in Figure [3](#fig:sector_dist), GPT-5-mini performed markedly better in domains GPT-4o-mini had previously failed almost entirely. In many of these cases, accuracy gains were both statistically significant and of large magnitude, suggesting that GPT-5-mini was more robust in domains underrepresented or terminologically complex in training data.



**Figure 3:** Hallucination rates across sectors where GPT-4o-mini performed worst. Error bars show 95% confidence intervals; green deltas () denote rate differences, with asterisks marking significance ().

By contrast, in the 15 most heavily represented sectors (Figure [4](#fig:sector_questions)), improvements were smaller and rarely significant: Although GPT-5-mini outperformed GPT-4o-mini in 11 of 15 sectors, only one gain achieved statistical significance. Here, higher baseline accuracy, more standardized question phrasing, and larger sample sizes likely limited the headroom for measurable improvement.

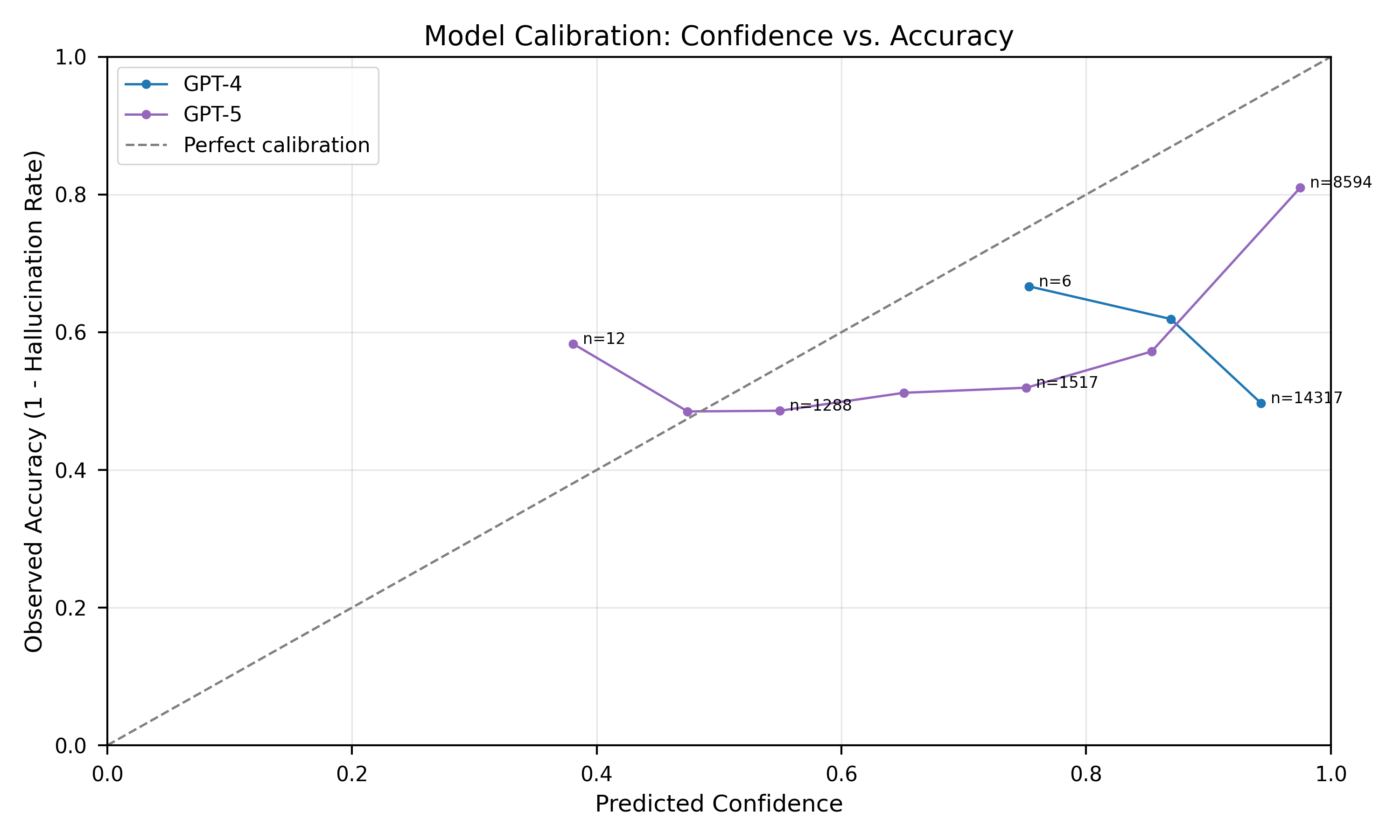


**Figure 4:** Hallucination rates for the 15 most heavily represented biomedical sectors in MedQuAD. Error bars show 95% confidence intervals; green deltas () denote rate differences, with asterisks marking significance ().

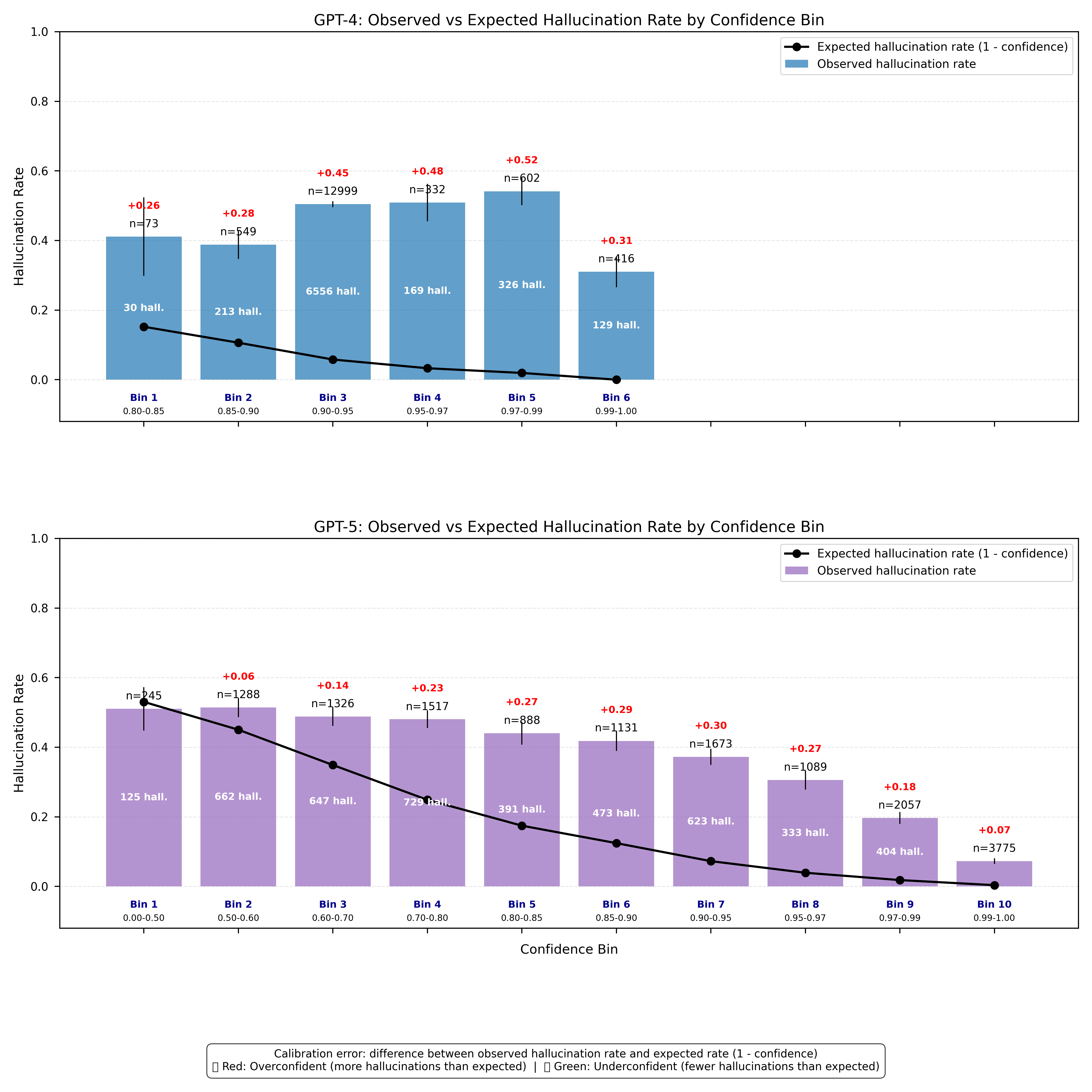
Together, Figures [3](#fig:sector_dist) and [4](#fig:sector_questions) show a “long-tail” dynamic: rare or underrepresented domains display the largest improvements, while common domains exhibit smaller changes. This distributional pattern indicates that improvements are uneven across sectors, with notable progress in harder cases but limited gains where baseline performance was already strong.

## 3.3 Confidence and Calibration

Next, we examined whether self-reported confidence aligned with factual accuracy. Figure [5](#fig:calibration) shows calibration curves comparing predicted confidence with observed accuracy. Both models were overconfident, with actual accuracy falling below reported confidence across most bins. GPT-5-mini, however, tracked the diagonal more closely than GPT-4o-mini, particularly at higher confidence levels, indicating improvement.

**Figure 5:** Calibration curves comparing predicted confidence with observed accuracy. GPT-5-mini displays closer alignment to perfect calibration than GPT-4o-mini, though both remain overconfident.

A more detailed bin-level breakdown (Figure [6](#fig:calibration_detailed)) highlights these differences in calibration. GPT-4o-mini was consistently miscalibrated, with observed hallucination rates exceeding expectations in nearly every bin and reaching as high as +0.52 in the 0.97–0.99 range. By contrast, GPT-5-mini displayed a parabolic pattern: calibration was relatively accurate at the extremes—low-confidence responses and the very highest confidence bin—but errors peaked in the mid-confidence range (0.70–0.95), where overconfidence remained between +0.23 and +0.30. This pattern indicates that GPT-5-mini achieved much better calibration overall than GPT-4o-mini, though misalignment persists in the middle of the confidence spectrum.



**Figure 6:** Observed vs. expected hallucination rates by confidence bin. GPT-5-mini shows reduced calibration error compared to GPT-4o-mini, but overconfidence persists in mid-to-high ranges.

Figure [7](#fig:confidence_extremes) highlights how the models behaved at the extremes of confidence. For GPT-4o-mini, very high-confidence answers (0.98) were unreliable, with nearly half (48%) turning out to be hallucinations. GPT-5-mini performed much better in the same setting, reducing that rate to 18%. The difference was also clear at the other end of the spectrum: GPT-4o-mini’s lowest-confidence responses were essentially a coin flip, while GPT-5-mini’s low-confidence answers were correct 70% of the time. Taken together, these patterns show that GPT-5-mini not only reduced the frequency of high-confidence failures but also made low-confidence signals more meaningful, even if calibration across the full range was still incomplete.

A graph of a pie chart

AI-generated content may be incorrect.

**Figure 7:** Extreme confidence cases. GPT-5-mini reduced the share of high-confidence hallucinations relative to GPT-4o-mini and produced low-confidence answers that more often aligned with correctness.

## 3.4 Confidence Alignment with External Entailment Probabilities

We also compared self-reported confidence with external entailment probabilities from the NLI classifier to see whether confidence reflected actual reliability. For GPT-4o-mini, alignment was essentially absent: scores were pushed toward the upper end regardless of correctness, and correlation with entailment probability was negligible (Pearson , Spearman ). In practice, this meant that a high confidence value provided little information about whether a response was correct or hallucinated.

A graph of a model

AI-generated content may be incorrect.

**Figure 8:** Relationship between self-reported confidence and external entailment probabilities. GPT-4o-mini showed little alignment, while GPT-5-mini displayed a strong positive trend.

GPT-5-mini showed a very different pattern. Confidence scores scaled positively with entailment probability, producing a moderate Pearson correlation () and a strong Spearman rank correlation (). Increases in confidence were generally mirrored by higher external validation, and the distribution of points revealed clearer separation between correct answers and hallucinations. Unlike GPT-4o-mini, which compressed both outcomes into the same high-confidence region, GPT-5-mini distinguished them more effectively—though a nontrivial number of overconfident hallucinations remained.

# 4 Discussion

This study set out to evaluate hallucination behavior in medical large language models (LLMs), focusing on the comparative performance of GPT-4o-mini and GPT-5-mini across domain accuracy, error correction, and confidence calibration. The results show clear generational advances, though important limitations remain. By situating these findings in relation to prior work, the discussion underscores both the significance of recent improvements and the continued risks of deploying such systems in clinical contexts.

A key finding was GPT-5-mini’s superior error correction. In head-to-head comparisons, the newer model resolved more than twice as many of GPT-4o-mini’s hallucinations as it introduced new ones, producing a net improvement of 18.5 percentage points. These gains were driven largely by reductions in direct contradictions, a category particularly concerning for clinical use because of its potential to mislead patients and providers. At the same time, the persistence of substantial overlap—where both models failed on the same questions—highlights that some categories of medical queries remain intrinsically difficult. This aligns with earlier analyses of ChatGPT’s medical performance, which found accuracy plateaus in areas with ambiguous phrasing or low training representation [2, 6].

Domain-level patterns add nuance to these aggregate results. GPT-5-mini’s strongest gains were concentrated in rare or terminologically complex fields such as neurology and kidney disorders, whereas common domains with higher baseline accuracy saw only marginal change. This “long-tail” distribution suggests that newer training approaches may have improved robustness in underrepresented areas but offered limited additional benefit where models were already competent. For medical deployment, this pattern is both encouraging and cautionary: it shows potential for closing gaps in hard-to-model domains, but it also suggests that future gains may plateau as training data coverage becomes saturated. Overcoming these limitations may require approaches beyond scaling—such as integrating structured biomedical knowledge or emphasizing reasoning improvements—to ensure continued progress across all domains.

Confidence and calibration analyses offer further evidence regarding the limitations of recent progress. Both models remained consistently overconfident, yet GPT-5-mini more closely followed the ideal calibration curve. The bin-level breakdown showed a parabolic pattern: calibration was fairly accurate at the lowest and highest confidence levels but broke down in the middle, where error rates spiked. This mid-range zone is particularly important in clinical practice, since these are the predictions most likely to trigger human review, yet they remained disproportionately unreliable. Prior studies have emphasized that such mismatches—where models present uncertain outputs with undue confidence—pose a serious risk to safe deployment [7, 8]. GPT-5-mini reduced high-confidence hallucinations compared to GPT-4o-mini, but the persistence of overconfidence, especially in the mid-range, highlights a limitation that newer model generations have yet to resolve.

Extreme cases also underscored both progress and ongoing risk. GPT-4o-mini hallucinated on nearly half of its responses at very high confidence (0.98), while GPT-5-mini cut this to 18%. At the other end, GPT-5-mini’s lowest-confidence outputs were correct 70% of the time, compared to GPT-4o-mini’s near-random performance. These shifts indicate that confidence scores now provide more reliable indicators at both extremes: low-confidence responses from GPT-5-mini more accurately marked uncertainty, while high-confidence answers were less likely than before to conceal hallucinations. Nonetheless, persistent high-confidence hallucinations underscore the need for external safeguards, including automated verification and structured human oversight.

Additional evidence of calibration improvement came from alignment with external entailment probabilities. GPT-5-mini showed moderate-to-strong positive correlations between confidence and factuality, whereas GPT-4o-mini displayed almost none. This suggests that GPT-5-mini’s confidence values better reflect factual reliability, providing further evidence of progress toward more interpretable and dependable calibration.

Overall, GPT-5-mini represents measurable but incomplete progress in reducing hallucinations and improving calibration. Gains were concentrated in rare biomedical sectors and in lowering the frequency of extreme-confidence errors, yet uneven performance and residual overconfidence remain. These findings demonstrate that even as models become more robust, reliance on their outputs—or on the confidence with which they are delivered—is still insufficient for ensuring reliability in high-stakes settings. Future work should test generalizability across datasets, develop calibration methods tailored to medical contexts, and explore hybrid systems that combine LLMs with structured biomedical knowledge.

# 5 Conclusion

This study compared GPT-4o-mini and GPT-5-mini on medical question–answering, focusing on hallucination rates, domain-level performance, and confidence calibration. GPT-5-mini showed clear gains, correcting many of GPT-4o-mini’s errors, improving robustness in underrepresented biomedical sectors, and reducing extreme-confidence failures. However, uneven progress across domains and persistent overconfidence highlight that reliability remains incomplete. In sum, GPT-5-mini shows selective improvements toward trustworthy medical LLMs, underscoring the need for continued calibration research and external safeguards before clinical deployment.

# Acknowledgements

The authors received no specific funding and have no acknowledgements to declare.

# References

1. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang Y, Madotto A, Fung P. Survey of hallucination in natural language generation. ACM Comput Surv. 2023;55(12):1-38. doi:10.1145/3571730

2. Rao A, Pang T, Kim J, et al. Assessing the factual accuracy of ChatGPT for medical information. JAMA Netw Open. 2023;6(8):e2325747. doi:10.1001/jamanetworkopen.2023.25747

3. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education. PLOS Digit Health. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198

4. Kadavath S, Conerly T, Ramesh A, et al. Language models (mostly) know what they know. arXiv [Preprint]. 2022. Available from: https://arxiv.org/abs/2207.05221

5. Si S, Wang R, Yu Y, et al. Measuring and improving calibration of large language models. In: Advances in Neural Information Processing Systems. 2023. Available from: https://openreview.net/forum?id=tgRkbY1iM5

6. Meng X, et al. The application of large language models in medicine. J Biomed Inform. 2024.

7. He P, Gao J, Chen W. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv [Preprint]. 2021. Available from: https://arxiv.org/abs/2006.03654

8. Lee P, et al. Clinical evaluation of large language models across accuracy, hallucinations, specificity, empathy, and actionability. npj Digit Med. 2025;8(1):18. doi:10.1038/s41746-025-01830-9

9. Laurer M. DeBERTa-v3-large-mnli-fever-anli-ling-wanli. Hugging Face. 2022. Available from: https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

10. Afroz. MedQuAD: Medical Question–Answer Dataset. Kaggle. 2024. Available from: https://www.kaggle.com/datasets/pythonafroz/medquad-medical-question-answer-for-ai-research

# 

# Appendix A: Prompt Templates

Two standardized prompts were used for consistency across models. They were designed to (1) maximize response consistency, (2) enforce concision and evidence-based language, and (3) enable reproducible confidence calibration analysis.

## A.1 Response Gathering Prompt

"role": "system", "content": "You are a specialized medical expert focusing on precise, evidence-based responses."  
You are a highly knowledgeable medical expert providing concise, accurate information. Answer the following medical question in a single, focused paragraph. Give a complete answer that covers all important aspects of the specific question asked, while being as concise as possible. Use proper medical terminology but keep the response clear and accessible. Every word should add value - include all crucial information related to the question, but avoid straying beyond what was asked. Keep responses under 100 words while maximizing information density.  
Question: {question}

## A.2 Confidence Rating Prompt

Rate your confidence in the accuracy and completeness of this response as a precise number between 0.00 and 1.00. Consider subtle differences in accuracy and completeness when choosing between similar numbers (like 0.87 vs 0.86). Be exact in your assessment - each hundredth decimal point matters.  
Return only the number with exactly two decimal places, nothing else.  
Question: {question} Response: {response}