Uncertainty-Augmented Prompting Reduces Unsafe Completions in Medical LLMs Facing Counterfactual Evidence

Anonymous Authors

Abstract

Large language models deployed in medical settings answer confidently regardless of input quality, posing patient safety risks when presented with counterfactual or implausible evidence. We investigate whether UA-PROMPT—a prompt-based intervention that instructs models to assess evidence quality before answering—can reduce unsafe completions on counterfactual medical questions. Using two frontier models (GPT-4.1 and CLAUDE SONNET 4.5) across three medical datasets (MEDQA, MED-HALT FAKE, and MED-HALT FCT) totaling 1,200 API calls, we compare baseline prompting against uncertainty-augmented prompting. Our central finding is that UA-PROMPT dramatically reduces unsafe completions on counterfactual questions—from 97% to 9% for GPT-4.1 and from 71% to 2% for CLAUDE SONNET 4.5 ($p < 0.0001$, Cohen's $d = 2.3$–4.2). Self-reported confidence under UA-PROMPT achieves an AUROC of 0.88–0.89 for counterfactual detection, substantially exceeding the prior best medical uncertainty estimation result of 0.58. Critically, the accuracy–safety tradeoff is model-dependent: GPT-4.1 preserves accuracy perfectly (92% → 92%) while CLAUDE SONNET 4.5 shows degradation partly attributable to structured output compliance failures. The intervention is specific to implausible premises and does not improve detection of subtle factual errors. Our results demonstrate that a zero-cost prompting intervention can activate latent evidence-assessment capabilities in frontier LLMs, providing an immediately deployable safety layer for medical applications.

1 Introduction

Frontier large language models now surpass the passing threshold on medical licensing examinations [Ness et al., 2024] and are increasingly integrated into clinical decision support. Yet these models share a dangerous default behavior: they answer confidently regardless of whether the input is medically sound. When presented with questions built on fabricated diseases, impossible pharmacology, or absurd physiological claims, current models engage with the false premise and produce fluent, confident—but unsafe—responses. In our experiments, GPT-4.1 confidently answered 97% of counterfactual medical questions without hesitation under standard prompting.

This overconfidence is not merely an academic concern. A model that explains treatment options for a fictional disease, or recommends dosing based on fabricated pharmacokinetics, poses a direct patient safety risk if its output influences clinical decisions. The practical question motivating this work is straightforward: **can we teach LLMs to say "I'm not sure" when the medical evidence looks implausible or unsafe?**

Three research streams address parts of this problem in isolation. First, uncertainty estimation methods—including semantic entropy [Kuhn et al., 2023, Farquhar et al., 2024], verbalized confidence [Tian et al., 2023, Xiong et al., 2023], and conformal prediction [Wang et al., 2024]—achieve AUROC 0.75–0.85 on general-domain QA but collapse to near-random performance on medical

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

benchmarks [Wu et al., 2024]. Second, adversarial robustness testing such as MedFuzz [Ness et al., 2024] demonstrates accuracy drops under perturbation but does not measure uncertainty. Third, counterfactual resistance methods such as AFICE [Zhao et al., 2025] train models to resist opposing arguments through DPO alignment but have not been tested in medical contexts. **No existing work combines uncertainty-augmented prompting with counterfactual medical inputs to measure whether LLMs become appropriately cautious.**

We fill this gap with a simple, deployment-ready intervention: UA-PROMPT, a prompt-engineering approach that explicitly instructs models to assess evidence quality before answering and to express uncertainty when premises appear implausible. We evaluate UA-PROMPT on two frontier models across three medical datasets, comparing baseline prompting against uncertainty-augmented prompting across 1,200 API calls.

Our results are striking. UA-PROMPT reduces unsafe completions on counterfactual medical questions from 97% to 9% for GPT-4.1 and from 71% to 2% for CLAUDE SONNET 4.5, with overwhelming statistical significance ($p < 0.0001$) and very large effect sizes (Cohen's $d = 2.3$–4.2). Self-reported confidence under UA-PROMPT achieves AUROC 0.88–0.89 for distinguishing counterfactual from legitimate questions, substantially exceeding the prior best medical uncertainty estimation result of 0.58 [Wu et al., 2024]. GPT-4.1 preserves accuracy perfectly (92% → 92%), demonstrating that the safety–accuracy tradeoff is model-dependent rather than inherent.

In summary, our main contributions are:

- We propose UA-PROMPT, a zero-cost prompting intervention that instructs LLMs to assess evidence quality before answering medical questions, reducing unsafe completions on counterfactual inputs by 88 and 69 percentage points for GPT-4.1 and CLAUDE SONNET 4.5, respectively.
- We conduct the first systematic evaluation combining uncertainty-augmented prompting with counterfactual medical inputs, testing 1,200 model responses across three medical datasets and two frontier models.
- We demonstrate that self-reported confidence under UA-PROMPT achieves AUROC 0.88–0.89 for counterfactual detection, substantially exceeding prior medical uncertainty estimation methods, and that the accuracy–safety tradeoff depends on model-specific instruction-following characteristics rather than being inherent to the intervention.

2 Related Work

We organize prior work along three axes: uncertainty estimation methods for LLMs, medical-domain challenges, and adversarial robustness with counterfactual resistance.

Uncertainty estimation for LLMs. Token-level methods such as predictive entropy and perplexity are computationally efficient but conflate linguistic uncertainty (multiple valid phrasings) with semantic uncertainty (genuine doubt about correctness) [Wu et al., 2024]. Kuhn et al. [2023] introduced semantic entropy to address this confusion by computing entropy over meaning clusters identified via natural language inference, achieving AUROC 0.75–0.85 on general QA tasks. Farquhar et al. [2024] extended this work to a broader range of models in a Nature publication, confirming strong confabulation detection but noting that the method cannot detect systematic errors where the model consistently produces the same wrong answer. Wang et al. [2025] proposed word-sequence entropy with semantic relevance weighting, improving performance on medical free-form QA. Verbalized confidence—asking models to self-report certainty—shows promise but tends toward overconfidence and is sensitive to prompt design [Kadavath et al., 2022, Lin et al., 2022, Tian et al., 2023, Xiong et al., 2023]. Wang et al. [2024] applied conformal prediction to LLMs for coverage-guaranteed abstention. Unlike these methods, which require multiple samples, probe training, or logit access, UA-PROMPT operates in a single forward pass with any black-box API.

Uncertainty estimation in the medical domain. Wu et al. [2024] provide the most comprehensive evaluation of uncertainty estimation methods in medicine, testing token probability, verbalized confidence, consistency-based, and self-verification approaches across MedQA, MedMCQA, PubMedQA, and COVID-QA. Their key finding is that methods achieving AUROC 0.75+ on general-domain QA collapse to near-random performance ($\sim$0.50) on medical benchmarks, with the best method (Two-Phase Verification) achieving only 0.58. This collapse reflects the difficulty of medical reasoning, where multi-step inference, precise terminology, and conditional dependencies challenge uncertainty estimation. Umapathi et al. [2023] introduced the MED-HALT benchmark for evaluating hallucination in medical LLMs, providing both reasoning hallucination tests with counterfactual premises and fact-checking tasks. Our work builds directly on this benchmark but uses it to evaluate prompt-based uncertainty interventions rather than measuring raw hallucination rates.

Adversarial robustness and counterfactual resistance. Ness et al. [2024] introduced MedFuzz, an adversarial fuzzing framework that reduces GPT-4 accuracy from 90.2% to 85.4% through subtle perturbations to medical questions. Critically, MedFuzz measures only accuracy degradation, not whether models' uncertainty signals appropriately increase—the gap our work addresses. Zhao et al. [2025] proposed AFICE with bilateral confidence estimation and DPO-based alignment to resist opposing arguments, demonstrating that models can be trained to maintain correct answers despite persuasive counterarguments. However, AFICE requires fine-tuning and has not been evaluated in medical contexts. Our work is complementary: we demonstrate that a zero-cost prompting intervention can achieve large reductions in unsafe completions without any model modification, though fine-tuning approaches like AFICE may provide additional benefits. Zeng et al. [2024] showed that uncertainty estimates in LLMs are fragile to semantically irrelevant input changes, raising concerns about deploying uncertainty-based safety mechanisms. Our results provide a counterpoint: for detecting grossly implausible premises, prompt-based uncertainty proves highly effective despite potential fragility on more subtle perturbations.

3 Methodology

We evaluate whether prompting LLMs to assess evidence quality before answering reduces unsafe completions on counterfactual medical inputs. Our experimental design compares two prompting conditions (baseline vs. uncertainty-augmented) across two frontier models and three medical datasets.

3.1 Models

We evaluate two frontier LLMs with strong medical QA capabilities:

- GPT-4.1 (OpenAI, direct API access): a frontier model with strong instruction following and structured output compliance.
- CLAUDE SONNET 4.5 (Anthropic, via OpenRouter API): a frontier model with strong reasoning capabilities.

Both models are accessed via OpenAI-compatible APIs at temperature 0.3 with a maximum of 500 output tokens. We use low temperature to improve reproducibility across runs.

3.2 Datasets

We sample 100 questions from each of three medical datasets (seed = 42), yielding 300 evaluation items per model per condition:

MEDQA [Jin et al., 2021] consists of USMLE-style multiple-choice questions drawn from medical board examinations. These are legitimate medical questions that serve as our control condition for measuring baseline accuracy and confidence calibration.

MED-HALT FAKE [Umapathi et al., 2023] contains absurd or counterfactual medical questions with fabricated premises (e.g., questions about diseases that do not exist or treatments based on impossible pharmacology). This is our primary test condition: a safe model should flag these questions as implausible rather than engaging with the false premise.

MED-HALT FCT [Umapathi et al., 2023] presents questions where a student's answer is incorrect and the model must identify the correct answer. Unlike MED-HALT FAKE, the premises in MED-HALT FCT are medically sound; the challenge is identifying factual errors in proposed answers.

3.3 Prompting Conditions

Baseline prompt. The model receives a medical expert persona and is asked to answer the question in structured JSON format, reporting an `answer_index`, `answer_text`, `confidence` (1–10 scale), and `reasoning`.

Metric	Type	Definition
Cautiousness rate	Safety	Fraction of responses where the model selects “I don’t know” or reports confidence ≤ 3
Unsafe completion rate	Safety	Fraction of responses with confidence ≥ 7 and an incorrect answer (or confident non-IDK on counterfactual inputs)
Accuracy	Performance	Fraction of correct answers on MEDQA and MED-HALT FCT
AUROC	Calibration	Area under the ROC curve using inverted self-reported confidence to discriminate legitimate vs. counterfactual questions
Evidence flag rate	Detection	Fraction of UA-PROMPT responses flagging evidence as suspicious, implausible, or dangerous

Table 1: Evaluation metrics. Safety metrics capture the model’s willingness to express uncertainty on suspicious inputs. Performance metrics verify that the intervention does not degrade medical knowledge.

Uncertainty-augmented prompt (UA-PROMPT). The model receives an extended persona that explicitly instructs it to: (1) assess evidence quality by checking for implausible, absurd, counterfactual, or dangerous claims before answering; (2) express appropriate uncertainty by selecting “I do not know” when available and assigning low confidence (1–3) on suspicious inputs; (3) answer confidently only when the question is medically sound. The model additionally reports an `evidence_quality` field with values: `normal`, `suspicious`, `implausible`, or `dangerous`.

Both conditions request identical structured JSON output, enabling direct comparison of answer selection, confidence, and reasoning across conditions.

3.4 Evaluation Metrics

We evaluate five primary metrics:

3.5 Statistical Analysis

We use chi-squared tests for comparing proportions (cautiousness rate, unsafe rate) between conditions and independent t-tests for comparing mean confidence scores. We report Cohen’s d for confidence effect sizes and odds ratios for cautiousness rate changes. All tests use significance level $\alpha = 0.05$.

3.6 Hypotheses

We pre-specify four hypotheses:

H1 (Cautiousness): UA-PROMPT increases the rate of cautious responses on counterfactual questions.

H2 (Safety): UA-PROMPT reduces the rate of confident, incorrect answers on counterfactual inputs.

H3 (Calibration): UA-PROMPT improves AUROC for distinguishing counterfactual from legitimate questions via self-reported confidence.

H4 (Accuracy preservation): UA-PROMPT does not significantly reduce accuracy on legitimate medical questions (threshold: $\leq$5 percentage points).

3.7 Implementation Details

We execute 1,200 API calls total (2 models × 2 conditions × 3 datasets × 100 samples). Responses are cached using SHA-256 hashing to enable reproducible re-analysis. We implement exponential backoff with a maximum of 5 retries per call. The overall parse success rate is 94.0%, with per-condition breakdowns reported in section 4.6.

Model	Condition	Cautiousness	Unsafe Rate	Mean Conf. (± std)	Evidence Flagged
GPT-4.1	Baseline	3.0%	97.0%	9.5 ± 1.0	—
	UA-PROMPT	**91.0%**	**9.0%**	3.5 ± 3.5	93.0%
CLAUDE SONNET 4.5	Baseline	22.0%	71.0%	8.0 ± 1.8	—
	UA-PROMPT	**88.0%**	**2.0%**	1.8 ± 1.1	90.0%

Table 2: Results on MED-HALT FAKE (counterfactual questions). UA-PROMPT reduces unsafe completion rates from 97% to 9% (GPT-4.1) and from 71% to 2% (CLAUDE SONNET 4.5). All differences are significant at $p < 0.0001$. Best results in **bold**.

Model	Condition	AUROC	N (legit / counterfactual)
GPT-4.1	Baseline	0.667	100 / 100
	UA-PROMPT	**0.879**	100 / 100
CLAUDE SONNET 4.5	Baseline	0.660	99 / 94
	UA-PROMPT	**0.892**	97 / 90

Table 3: AUROC for counterfactual detection using inverted self-reported confidence. UA-PROMPT improves AUROC from ∼0.66 to ∼0.89, exceeding the prior best medical uncertainty estimation result of 0.58 [Wu et al., 2024]. Best results in **bold**.

4 Results

4.1 Unsafe Completions on Counterfactual Questions (H1, H2)

Table 2 presents our primary finding: UA-PROMPT dramatically reduces unsafe completions on counterfactual medical questions for both models.

Under baseline prompting, GPT-4.1 confidently engages with absurd medical premises in 97% of cases, with a mean confidence of 9.5/10. CLAUDE SONNET 4.5 shows marginally more caution at baseline (22% cautiousness), suggesting some inherent resistance to counterfactual inputs. Under UA-PROMPT, both models shift to predominantly cautious behavior: cautiousness rises to 91% and 88%, while unsafe rates drop to 9% and 2%, respectively. Both models successfully flag evidence quality in ∼90% of uncertainty-condition responses.

All counterfactual results are statistically significant ($p < 0.0001$, chi-squared tests). The effect sizes are extraordinarily large: the odds ratio for GPT-4.1 cautiousness is 326.9, and for CLAUDE SONNET 4.5 is 26.0. Cohen's d for the confidence shift is 2.31 (GPT-4.1) and 4.19 (CLAUDE SONNET 4.5), well above the conventional threshold of 0.8 for "large" effects.

4.2 Counterfactual Detection via Confidence (H3)

Table 3 shows that self-reported confidence under UA-PROMPT is an effective counterfactual detector.

The baseline AUROC of ∼0.66 indicates that even without explicit uncertainty instructions, both models assign slightly lower confidence to counterfactual questions—but not nearly enough to serve as a reliable safety mechanism. Under UA-PROMPT, AUROC increases to 0.88–0.89, creating clear separation between confidence distributions on legitimate and counterfactual inputs.

Comparison with prior work. Wu et al. [2024] reported the best prior AUROC for medical uncertainty estimation at 0.58 (Two-Phase Verification method on general medical QA). Our UA-PROMPT achieves 0.88–0.89, a substantial improvement. However, a direct comparison requires caution: our task (distinguishing fabricated questions from real ones) involves coarser discrimination than identifying correct vs. incorrect answers on legitimate medical questions. The high AUROC reflects the effectiveness of the intervention for a specific safety use case rather than a general advance in medical uncertainty estimation.

Model	Condition	MedQA Accuracy	Med-HALT FCT Accuracy
GPT-4.1	Baseline	92.0%	89.0%
	UA-Prompt	**92.0%**	**88.0%**
Claude Sonnet 4.5	Baseline	93.0%	82.0%
	UA-Prompt	66.0% ↓	75.0% ↓

Table 4: Accuracy on legitimate medical questions. GPT-4.1 preserves accuracy perfectly under UA-Prompt. Claude Sonnet 4.5 shows significant degradation, partly attributable to a 34% parse failure rate under the uncertainty prompt. Best results in **bold**.

Model	Condition	Cautiousness	Unsafe Rate	Mean Conf.	Accuracy
GPT-4.1	Baseline	0.0%	11.0%	9.7	89.0%
	UA-Prompt	1.0%	12.0%	9.3	88.0%
Claude Sonnet 4.5	Baseline	3.0%	15.0%	8.8	82.0%
	UA-Prompt	9.0%	18.0%	8.4	75.0%

Table 5: Results on Med-HALT FCT (fact-checking). No significant differences between conditions for either model (all $p > 0.05$). The intervention targets implausible premises, not subtle factual errors.

4.3 Accuracy Preservation (H4)

Table 4 reveals that the accuracy–safety tradeoff is model-dependent. GPT-4.1 preserves accuracy perfectly: 92% → 92% on MedQA and 89% → 88% on Med-HALT FCT. The uncertainty prompt does not cause GPT-4.1 to second-guess valid medical knowledge.

Claude Sonnet 4.5 shows a significant accuracy drop: 93% → 66% on MedQA ($p = 0.001$), exceeding our pre-specified 5% degradation threshold. However, this drop is substantially confounded by a **34% parse failure rate** on MedQA under the uncertainty prompt (vs. 4% at baseline; see section 4.6). When Claude Sonnet 4.5 receives the more complex uncertainty prompt, it frequently produces verbose narrative responses instead of the requested JSON format, and parse failures are counted as incorrect. GPT-4.1 maintains 0% parse error across all conditions.

4.4 Fact-Checking Dataset: Limited Impact

On Med-HALT FCT, uncertainty prompting shows no meaningful improvement for either model (table 5).

None of the Med-HALT FCT differences reach statistical significance. This result is interpretable: Med-HALT FCT questions present real medical questions where a student gave a wrong answer—the premise is legitimate, and the task requires medical knowledge rather than evidence-quality assessment. UA-Prompt is designed to detect implausible premises, not subtle factual errors, and both models appropriately maintain high confidence on these medically sound questions.

4.5 Confidence Distributions

The confidence distributions reveal the mechanism of the intervention clearly. Under baseline prompting, both models cluster at high confidence (8–10) across all datasets, including counterfactual questions. GPT-4.1 is especially extreme, with mean confidence of $9.96/10$ on MedQA and $9.47/10$ even on counterfactual questions.

Under UA-Prompt on counterfactual questions, confidence distributions shift dramatically downward (mean 1.8 for Claude Sonnet 4.5, 3.5 for GPT-4.1), creating clear bimodal separation from legitimate questions. On legitimate and fact-checking questions, confidence remains high (8.0–9.6), confirming that the prompt does not indiscriminately suppress confidence.

Model	Condition	MedQA	Med-HALT Fake	Med-HALT FCT
GPT-4.1	Baseline	100%	100%	100%
	UA-Prompt	100%	100%	100%
Claude Sonnet 4.5	Baseline	96%	91%	97%
	UA-Prompt	**66%**	90%	88%

Table 6: JSON parse success rates. GPT-4.1 achieves perfect compliance. Claude Sonnet 4.5 shows degraded compliance under UA-Prompt, particularly on MedQA (66%), partly explaining the observed accuracy drop.

Hypothesis	GPT-4.1	Claude Sonnet 4.5	Effect Size	Verdict
H1 (Cautiousness)	3% → 91%, $p < 0.0001$	22% → 88%, $p < 0.0001$	OR = 327 / 26	**Supported**
H2 (Safety)	97% → 9%, $p < 0.0001$	71% → 2%, $p < 0.0001$	—	**Supported**
H3 (Calibration)	0.667 → 0.879	0.660 → 0.892	—	**Supported**
H4 (Accuracy)	92% → 92%	93% → 66%	—	**Mixed**

Table 7: Summary of hypothesis testing. H1–H3 are strongly supported. H4 is model-dependent: GPT-4.1 preserves accuracy while Claude Sonnet 4.5 shows degradation partly attributable to parse errors.

4.6 Parse Success Rates

Table 6 reveals an important practical finding: more complex system prompts with multi-step reasoning instructions can degrade structured output compliance in some models. Claude Sonnet 4.5's 34% parse failure rate on MedQA under UA-Prompt indicates that the model struggles with competing generation pressures—assessing evidence quality while also producing structured JSON—particularly on legitimate questions where it has a strong answer.

4.7 Hypothesis Testing Summary

5 Discussion

5.1 Prompt-Based Safety Guardrails Activate Latent Capabilities

Our central finding is both simple and powerful: explicitly asking LLMs to assess evidence quality before answering causes them to correctly flag implausible medical questions ~90% of the time, compared to ~3–22% at baseline. This is a zero-cost intervention requiring no fine-tuning, no additional infrastructure, and no access to model internals.

The mechanism is clear from the data. The uncertainty prompt activates the model's *existing* ability to recognize implausible premises—an ability that is suppressed by standard task-focused prompts. Both models already "know" that counterfactual questions are absurd, as evidenced by their slightly lower baseline confidence on these inputs (AUROC ~0.66). However, they default to answering confidently because standard prompts ask for an answer, not an evidence assessment. UA-Prompt redirects the model's attention to evidence quality before answer generation, and this reordering is sufficient to produce dramatically different behavior.

5.2 Specificity: Implausible Premises vs. Subtle Errors

The intervention works specifically for obviously implausible premises (Med-HALT Fake) but not for subtle factual errors (Med-HALT FCT). This specificity is both a strength and a limitation. As a strength, it means the intervention provides a reliable guardrail against grossly implausible inputs—questions about fictional diseases, impossible physiological claims, or fabricated pharmacology. As a limitation, it should not be expected to improve performance on tasks requiring nuanced medical reasoning, where the premises are sound but the correct answer requires careful knowledge application.

Method	**Context**	**AUROC**	**Access Required**
Token probability methods [Wu et al., 2024]	Medical QA, correct vs. incorrect	∼0.50	Logits
Two-Phase Verification [Wu et al., 2024]	Medical QA, correct vs. incorrect	0.58	Black-box
Semantic entropy [Farquhar et al., 2024]	General QA, confabulation detection	0.75–0.85	Logits
UA-PROMPT (baseline)	Medical, legit vs. counterfactual	0.66	Black-box
UA-PROMPT (uncertainty)	Medical, legit vs. counterfactual	**0.88–0.89**	Black-box

Table 8: Comparison with prior uncertainty estimation methods. UA-PROMPT achieves the highest AUROC in a medical context, though the task (distinguishing fabricated from real questions) involves coarser discrimination than prior work. Best result in **bold**.

5.3 The Accuracy–Safety Tradeoff Is Model-Dependent

The divergent accuracy results between GPT-4.1 (no degradation) and CLAUDE SONNET 4.5 (significant degradation) reveal that the accuracy–safety tradeoff is not inherent to the intervention but depends on model characteristics.

GPT-4.1 exhibits strong instruction following: it applies the uncertainty assessment to evidence quality as instructed, determines that legitimate MEDQA questions are sound, and proceeds to answer with high confidence and maintained accuracy. It perfectly compartmentalizes the two tasks (evidence assessment and answer selection).

CLAUDE SONNET 4.5 struggles with the multi-faceted prompt, producing verbose narrative responses instead of structured JSON on legitimate questions (34% parse failure rate). This suggests that the more complex prompt creates competing generation pressures that degrade output format compliance, which manifests as apparent accuracy loss.

This finding has a clear practical implication: uncertainty-augmented prompts should be paired with structured output constraints (e.g., JSON mode, function calling) to prevent format compliance degradation. Models with weaker instruction following may require simpler prompt formulations or output enforcement mechanisms.

5.4 Comparison with Prior Work

Table 8 contextualizes our AUROC results. Our results exceed prior medical uncertainty estimation methods by a wide margin, but the comparison is not fully apples-to-apples. Our task (distinguishing fabricated questions from real ones) is arguably a coarser discrimination than identifying which answers to real questions are correct vs. incorrect. The high AUROC reflects the success of the intervention for a specific safety use case rather than a general advance in medical uncertainty estimation.

5.5 Limitations

Sample size. We evaluate 100 questions per dataset, sufficient for detecting the large effects observed but potentially underpowered for detecting smaller effects. The non-significant MED-HALT FCT results could reflect either a true null effect or insufficient power.

Dataset representativeness. MED-HALT FAKE questions may be more obviously absurd than real-world medical misinformation. Sophisticated misinformation with plausible but subtly wrong premises would be harder to detect and represents an important direction for future evaluation.

Single-turn evaluation. We test single questions in isolation. In real clinical workflows, counterfactual information may be embedded in longer patient histories or accumulate gradually through multi-turn conversations.

Prompt sensitivity. We test one uncertainty prompt formulation. The optimal design likely varies by model and may require task-specific tuning. The CLAUDE SONNET 4.5 format compliance issue highlights that prompt robustness is a practical concern.

Parse error confound. CLAUDE SONNET 4.5's apparent accuracy drop is partially attributable to format compliance failures rather than genuine knowledge degradation. A fairer evaluation would use structured output constraints (e.g., function calling).

No fine-tuning comparison. We compare only prompt-based interventions. Fine-tuned approaches such as AFICE-style DPO alignment [Zhao et al., 2025] may achieve better accuracy preservation alongside uncertainty awareness.

6 Conclusion

We introduce UA-PROMPT, a prompt-based intervention that instructs LLMs to assess evidence quality before answering medical questions. Our experiments on two frontier models across 1,200 API calls demonstrate five key findings: (1) UA-PROMPT reduces unsafe completions on counterfactual medical questions from 97% to 9% for GPT-4.1 and from 71% to 2% for CLAUDE SONNET 4.5 ($p < 0.0001$, Cohen's $d = 2.3$–4.2); (2) self-reported confidence under UA-PROMPT achieves AUROC 0.88–0.89 for counterfactual detection, substantially exceeding the prior best medical uncertainty estimation result of 0.58; (3) the accuracy–safety tradeoff is model-dependent, with GPT-4.1 preserving accuracy perfectly while CLAUDE SONNET 4.5 shows degradation partly due to structured output compliance failures; (4) the intervention is specific to implausible premises and does not improve detection of subtle factual errors; (5) even baseline frontier models are dangerously overconfident on counterfactual inputs, with GPT-4.1 confidently answering 97% of absurd medical questions.

These results suggest that frontier LLMs already possess the ability to recognize implausible medical evidence, but this ability is suppressed by standard task-focused prompts. A simple, zero-cost prompting intervention can activate this latent capability, providing an immediately deployable safety layer for medical applications.

Future work. Three directions emerge from our findings. First, evaluating UA-PROMPT on more sophisticated medical misinformation—plausible but subtly wrong premises—would test the boundaries of prompt-based safety. Second, combining prompt-based uncertainty with structured output constraints and fine-tuning approaches may address the format compliance issues observed with CLAUDE SONNET 4.5. Third, testing in multi-turn conversational settings where misinformation accumulates gradually would better approximate real clinical deployment conditions.

References

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Saurav Kadavath et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. In *Transactions on Machine Learning Research (TMLR)*, 2022.

Robert Ness et al. MedFuzz: Exploring the robustness of large language models in medical question answering. *Microsoft Research Technical Report*, 2024.

Katherine Tian et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models including large language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. Med-HALT: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, 2023.

Jinhao Wang et al. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

Zhiyuan Wang et al. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*, 150, 2025.

Yijie Wu et al. Uncertainty estimation of large language models in medical question answering. *arXiv preprint arXiv:2407.08662*, 2024.

Miao Xiong et al. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. *arXiv preprint arXiv:2306.13063*, 2023.

Zhisheng Zeng et al. The uncertainty in LLMs is fragile. *arXiv preprint arXiv:2407.15729*, 2024.

Liang Zhao et al. AFICE: Aligning for faithful integrity with confidence estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

A Full Results

Table 9 presents the complete results across all models, datasets, and conditions.

Model	Dataset	Condition	Caut.	Unsafe	Conf. (mean ± std)	Acc.	Parse	Ev. Flag
GPT-4.1	MedQA	Baseline	0.0%	8.0%	10.0 ± 0.2	92.0%	100%	—
GPT-4.1	MedQA	UA-Prompt	1.0%	7.0%	9.6 ± 0.9	92.0%	100%	0.0%
GPT-4.1	Med-HALT Fake	Baseline	3.0%	97.0%	9.5 ± 1.0	—	100%	—
GPT-4.1	Med-HALT Fake	UA-Prompt	91.0%	9.0%	3.5 ± 3.5	—	100%	93.0%
GPT-4.1	Med-HALT FCT	Baseline	0.0%	11.0%	9.7 ± 0.5	89.0%	100%	—
GPT-4.1	Med-HALT FCT	UA-Prompt	1.0%	12.0%	9.3 ± 1.1	88.0%	100%	1.0%
Claude Sonnet 4.5	MedQA	Baseline	0.0%	6.0%	8.9 ± 0.7	93.0%	96%	—
Claude Sonnet 4.5	MedQA	UA-Prompt	12.0%	19.0%	8.1 ± 2.7	66.0%	66%	0.0%
Claude Sonnet 4.5	Med-HALT Fake	Baseline	22.0%	71.0%	8.0 ± 1.8	—	91%	—
Claude Sonnet 4.5	Med-HALT Fake	UA-Prompt	88.0%	2.0%	1.8 ± 1.1	—	90%	90.0%
Claude Sonnet 4.5	Med-HALT FCT	Baseline	3.0%	15.0%	8.8 ± 1.3	82.0%	97%	—
Claude Sonnet 4.5	Med-HALT FCT	UA-Prompt	9.0%	18.0%	8.4 ± 2.3	75.0%	88%	4.0%

Table 9: Complete results across all experimental conditions. Caut. = cautiousness rate; Unsafe = unsafe completion rate; Conf. = mean confidence; Acc. = accuracy; Parse = JSON parse success rate; Ev. Flag = evidence flagged as suspicious/implausible/dangerous.

Table 10 reports all statistical tests.

Model	Dataset	Test	Baseline	UA-Prompt	p-value	Effect Size
GPT-4.1	Med-HALT Fake	Cautiousness (χ^2)	3.0%	91.0%	< 0.0001	OR $= 326.9$
GPT-4.1	Med-HALT Fake	Unsafe rate (χ^2)	97.0%	9.0%	< 0.0001	—
GPT-4.1	Med-HALT Fake	Confidence (t-test)	9.5	3.5	< 0.0001	$d = 2.31$
GPT-4.1	MedQA	Confidence (t-test)	10.0	9.6	0.0005	$d = 0.50$
Claude Sonnet 4.5	Med-HALT Fake	Cautiousness (χ^2)	22.0%	88.0%	< 0.0001	OR $= 26.0$
Claude Sonnet 4.5	Med-HALT Fake	Unsafe rate (χ^2)	71.0%	2.0%	< 0.0001	—
Claude Sonnet 4.5	Med-HALT Fake	Confidence (t-test)	8.0	1.8	< 0.0001	$d = 4.19$
Claude Sonnet 4.5	MedQA	Cautiousness (χ^2)	0.0%	12.0%	0.0011	—
Claude Sonnet 4.5	MedQA	Unsafe rate (χ^2)	6.0%	19.0%	0.0103	—
Claude Sonnet 4.5	Med-HALT FCT	Cautiousness (χ^2)	3.0%	9.0%	0.137	OR $= 3.20$

Table 10: Statistical tests for all experimental comparisons. OR = odds ratio; d = Cohen's d.