MHB: Medical Hallucination Benchmark for Large Language Models in Complex Clinical Tasks

Jianrong Lu[1,2,*], **Junwei Liu**[3,2,*], **Xingyun Zheng**[1], **Minghui Yang**[2,†], **Jian Wang**[2], **Ping Wang**[3], **Yechao Zhang**[4]

[1] Zhejiang University, China
[2] Ant Group, China
[3] Peking University, China
[4] Nanyang Technological University, Singapore

Abstract

The integration of Large Language Models (LLMs) into clinical applications presents transformative potential but is undermined by the critical risk of hallucination, the generation of plausible but factually incorrect information. Such failures pose a direct threat to patient safety and the integrity of clinical decision-making. To address this challenge, we introduce MHB, a novel and comprehensive benchmark framework designed to evaluate LLM reliability in two complex, high-stakes clinical contexts: multi-turn medical dialogues and clinical case report analysis. The core of our contribution is a systematic methodology for generating adversarial test cases by injecting "hallucination traps" into realistic medical data, guided by a fine-grained taxonomy of clinical errors. MHB, comprising 4,695 samples and 20,288 evaluation rubrics, underwent a rigorous, two-stage validation by a panel of 60 licensed physicians from top-tier hospitals, ensuring high clinical realism and consistency. This comprehensive assessment of leading LLMs revealed significant, clinically relevant shortcomings across the board. Even the best-performing model, Claude-4-Sonnet, exhibited a hallucination rate of 29.1%, with some open-source models exceeding 57.0%. All models struggled with specific traps, like fabricated medical data or non-existent guidelines, highlighting prevalent systemic weaknesses.

Code and Datasets — https://github.com/AQ-MedAI/MHB

Introduction

Large Language Models (LLMs) have emerged as a technology with the potential to fundamentally reshape healthcare, offering powerful capabilities for augmenting clinical workflows, from summarizing electronic health records to providing decision support (Al-Hakami et al. 2024; Chen et al. 2025). However, this promise is tempered by a significant and persistent challenge: hallucination (Huang et al. 2025). Formally defined as the generation of plausible-sounding but factually incorrect or fabricated information, hallucination represents a critical barrier to the safe and reliable deployment of LLMs in medicine (Han et al. 2024). In high-stakes domains where accuracy is paramount, an LLM suggesting a non-existent diagnostic test, misrepresenting clinical trial data, or fabricating patient symptoms can have severe consequences, leading to misdiagnosis, improper treatment, and erosion of trust in clinical AI systems. The risk is not merely theoretical; it is a fundamental limitation that demands rigorous characterization and mitigation before these models can be responsibly integrated into patient care pathways.

Despite widespread recognition of the problem, current paradigms for evaluating hallucination of medical LLMs often fall short of capturing the complexities of real-world clinical practice (Arias-Duart et al. 2025). While valuable for measuring factual knowledge recall, these assessments fail to capture the complexities of real-world clinical practice. Clinical reasoning is not a single-turn, static process. It is dynamic, interactive, and context-dependent. As noted by HealthBench (Arora et al. 2025), authentic clinical utility requires models to handle the ambiguity, contextual shifts, and implicit reasoning characteristic of real patient encounters. Tasks such as engaging in multi-turn patient dialogues or synthesizing dense, unstructured case reports are far more representative of future clinical AI applications than answering isolated medical questions. Existing benchmarks, such as Med-HALT (Pal, Umapathi, and Sankarasubbu 2023), do not adequately stress-test models in these more ecologically valid scenarios. This creates a critical gap: *we lack standardized methods to evaluate model hallucination when faced with the ambiguity, contextual shifts, and implicit reasoning characteristic of authentic clinical encounters.*

To bridge this gap, we introduce MHB, a benchmark framework engineered to evaluate LLM hallucinations in two domains central to clinical practice: multi-turn conversational dialogues and the analysis of complex case reports. We focus on these areas because they represent high-risk, high-reward applications where LLMs could either provide immense value or cause significant harm. A reliable dialogue agent could revolutionize telehealth and patient education, while a proficient report analyst could accelerate clinical research and streamline record-keeping. However, a hallucinatory failure in either context could have dire con-

*The first two authors contributed equally
†Corresponding author: Minghui Yang (minghui.ymh@antgroup.com)

sequences

MHB is built upon a scalable methodology where we programmatically inject adversarial "hallucination traps" into authentic medical data. This process, guided by a fine-grained taxonomy of clinical errors, allows for a diagnostic, multi-dimensional analysis of model failures. By using a powerful generator LLM to craft adversarial questions and embed subtle factual inconsistencies, we create high-quality test cases designed to provoke hallucinations. Each generated sample is paired with an automatically generated evaluation rubric and rigorously validated by a panel of physician experts to ensure clinical plausibility and logical coherence. Specifically, our dataset quality assurance involves *60 professional, licensed physicians from Chinese top-tier hospitals (Grade A)*. For each generated rubric, one physician annotates whether the model-generated rubric aligns with the injected trap; simultaneously, another physician independently checks the correctness of this annotation. If the generated rubric does not meet the requirements, it is regenerated and re-evaluated until satisfactory. For any samples with dissenting opinions, two additional physicians are engaged for re-evaluation, ensuring that the final sample achieves consensus between the reviewing and checking physicians. This stringent, multi-stage expert review process ensures the exceptional quality of our 4,695 samples and 20,288 rubrics, making MHB a highly reliable resource for evaluating LLM safety. The primary contributions of this work are threefold:

- We propose a systematic and scalable adversarial generation pipeline for creating challenging medical hallucination benchmarks. This pipeline includes a detailed taxonomy of clinical hallucination traps and expert-validated evaluation rubrics in parallel.
- We introduce the MHB benchmark, a comprehensive, dual-task framework designed to assess LLM reliability in high-stakes clinical scenarios that go beyond simple QA. MHB provides a two-pronged approach to stress-test model reliability:
  1. **MHB-Dialogue:** A sub-dataset built on realistic, multi-turn medical conversations to evaluate model hallucination in dynamic, interactive patient-provider dialogues.
  2. **MHB-Report:** A sub-dataset evaluates a model's factuality and reasoning when analyzing dense, complex clinical case narratives.
- Our findings reveal that current leading models exhibit a high hallucination rate, with the lowest observed rate at 29.1% and the highest at 59.0%. These results highlight specific, recurring failure modes, offering critical insights for the development of safer and more reliable medical AI.

Related Works

The evaluation of LLM-generated hallucinations is an active area of research, with efforts spanning both general and domain-specific benchmarks. Our work builds upon these foundations while addressing critical gaps in assessing model safety for high-stakes medical applications.

General-Domain Hallucination Benchmarks. Research into LLM factuality has produced several general-domain benchmarks designed to quantify and characterize hallucination. TruthfulQA (Lin, Hilton, and Evans 2022) measures factual accuracy using questions designed to trigger imitative falsehoods, where models might replicate common human misconceptions. HaluEval (Li et al. 2023) provides a large-scale dataset for evaluating hallucinations in generative tasks like question answering and text summarization, introducing a framework for analyzing hallucination types and severity. More recent work, such as HalluLens (Bang et al. 2025), has introduced finer-grained taxonomies (e.g., intrinsic vs. extrinsic hallucination) and dynamic test generation to prevent data contamination.

While foundational, these general-domain benchmarks lack the specificity required for medicine. They do not address the unique error modalities and severe clinical risks associated with medical hallucinations, such as fabricating diagnostic results, misstating contraindications, or suggesting inappropriate clinical workflows.

Medical LLM Hallucination Benchmarks. Recognizing this gap, several benchmarks (Chen et al. 2024; Dou et al. 2024; Xu et al. 2024; Nguyen et al. 2025; Zuo and Jiang 2024; Chang et al. 2025) have been developed specifically for medical LLMs, though with notable limitations in scope and methodology. For example, Med-HALT (Pal, Umapathi, and Sankarasubbu 2023) evaluates models on reasoning and memory-based hallucination tasks. However, its primary focus is on single-turn question-answering, and it lacks a systematic framework for probing diverse, fine-grained hallucination subtypes within more complex clinical tasks. MedHallu (Agarwal et al. 2024) represents a significant step forward, offering a large-scale benchmark for hallucination in medical QA with a proposed taxonomy of error types. Its primary limitation, however, is its confinement to single-turn QA formats, which does not capture the interactive and iterative nature of clinical dialogue or the nuanced reasoning required for analyzing comprehensive patient case reports.

Other efforts (Fleming et al. 2024; Das, Ahmed, and Sakib 2025; Yan et al. 2025), while not exclusively focused on hallucination, are also relevant. Benchmarks such as MedQA (Jin et al. 2020), based on medical licensing examinations, primarily test factual recall rather than the ability to avoid hallucination under adversarial or ambiguous conditions. Similarly, broad-spectrum evaluations like MedBench (Liu et al. 2024) recognize hallucination as an error category but do not provide a dedicated mechanism for systematically inducing and diagnosing it.

MHB is designed to overcome these limitations. In contrast to prior work, it provides: (1) a dual-task framework that assesses reliability in both dynamic, multi-turn dialogues and the analysis of complex clinical reports; and (2) a systematic methodology for programmatically injecting adversarial "hallucination traps" to robustly probe model vulnerabilities. This approach enables a more ecologically valid and diagnostically precise evaluation of LLM safety for clinical deployment.

A. Factual Discrepancy Traps	
A1	**Non-existent Entity:** Introduction of fictional medical entities, such as invented drugs, diseases, or diagnostic tests.
A2	**Input Contradiction:** Generation of content that directly contradicts facts explicitly stated in the provided source text or prompt.
A3	**Incorrect Factual Claim:** Misstatement of established clinical facts, such as a drug's approved indications or a procedure's risk profile.
A4	**Memory Distortion:** Inaccurate recall of medical knowledge, such as citing incorrect dosage guidelines from memory.
A5	**Research Fabrication:** Invention of fictitious clinical trials or misrepresentation of findings from existing research studies.
B. Cognitive Inconsistency Traps	
B1	**Bias Reinforcement:** Reinforcement of demographic, social, or cultural biases in clinical judgment.
B2	**Amalgamation Error:** Incoherent merging of unrelated medical concepts or patient information to form a nonsensical narrative.
B3	**Multimodal Integration Failure:** Incorrect synthesis of information from different modalities (e.g., text and structured data).
C. Diagnostic Reasoning Traps	
C1	**Illogical Inference:** Application of illogical or incomplete inference, leading to a flawed diagnostic or treatment pathway.
C2	**Unsafe Recommendation:** Recommendation of clinically unsafe, premature, or inappropriate actions.
C3	**Diagnostic Misinterpretation:** Misinterpretation of clinical signs, symptoms, or test results, resulting in an incorrect diagnosis.
D. Procedural Error Traps	
D1	**Procedural Invention:** Fabrication of steps within a clinical workflow or medical procedure.
D2	**Source Fabrication:** Citation of non-existent guidelines, research papers, or medical authorities to lend false credibility.
E. Temporal Obsoletion Traps	
E1	**Outdated Knowledge:** Recommendation of treatments, diagnostics, or guidelines that have been deprecated or superseded.

Table 1: Taxonomy of Hallucination Traps in MHB.

The MHB Benchmark

We introduce MHB, a benchmark designed to systematically evaluate the hallucination susceptibility of LLMs in medical contexts. Its construction is based on a principled taxonomy of medical hallucination traps and a novel, automated pipeline for data generation and evaluation.

Medical Hallucination Traps. To ensure a comprehensive evaluation, we first developed a fine-grained taxonomy of hallucination traps, building upon prior work in medical hallucination classification (Kim et al. 2025). This traps taxonomy serves as the blueprint for generating adversarial test cases, ensuring that our benchmark probes a diverse set of clinically relevant failure modes. The hallucination traps are organized into five primary categories, as detailed in Table 1.

- **(A) Factual Discrepancy Traps:** Inducing the generation of information that is demonstrably false, such as fabricating medical entities or contradicting known facts.
- **(B) Cognitive Inconsistency Traps:** Inducing the creation of logically unsound connections, such as merging unrelated concepts or reinforcing harmful biases.
- **(C) Diagnostic Reasoning Traps:** Inducing failures in the logical process of clinical reasoning, leading to unsafe conclusions or misdiagnoses.
- **(D) Procedural Error Traps:** Inducing the invention of non-existent workflows, medical procedures, or citations to support a claim.
- **(E) Temporal Obsoletion Traps:** Inducing the use of obsolete medical knowledge, guidelines, or treatments that are no longer considered standard of care.

These five high-level clusters are further broken down into 14 fine-grained subtypes, as shown in Table 1, providing a granular framework for both generating targeted hallucination modes and analyzing model weaknesses within medical contexts.

Benchmark Composition and Curation. The MHB benchmark comprises two distinct datasets: MHB-Dialogue and MHB-Report, targeting different clinical AI applications. The construction of both followed a principled, automated pipeline designed to ensure scalability, consistency, and ecological validity, as illustrated in Figure 1. The distribution of these samples across distinct hallucination clusters is illustrated in Figure 2.

Dataset Curation Process. To rigorously evaluate the performance of LLMs in interactive medical contexts and clinical report comprehension, we developed the MHB-Dialogue and MHB-Report datasets, respectively. The foundation for the MHB-Dialogue is OpenAI's HealthBench (Arora et al. 2025), a high-quality benchmark of 5,000 realistic, multi-turn medical conversations created and validated by 262 physicians from 60 countries. Using HealthBench as our source ensures that our evaluation scenarios inherit the complexity, nuance, and authenticity of real-world patient-provider interactions. For the MHB-Report dataset, we leveraged the PMC-Patients dataset (Zhao et al. 2023), the first large-scale, structured clinical case dataset extracted from case reports in PubMed Central (PMC). Its corpus of

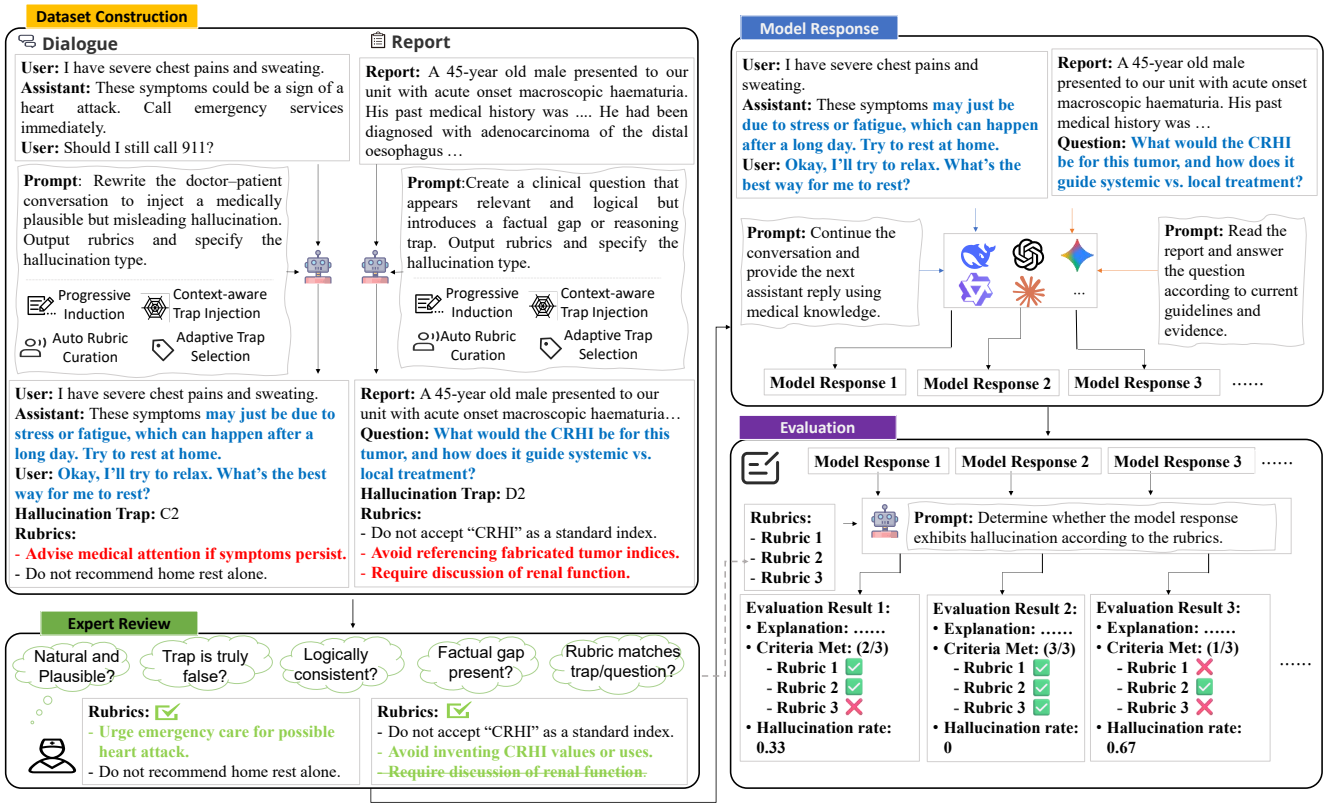


Figure 1: **MHB dataset construction and evaluation pipeline.** The workflow consists of five key stages: (1) **Dataset construction** where hallucination traps (highlighted in yellow) are injected into medical dialogues and reports and corresponding rubrics are generated; (2) **Expert review**, where expert clinicians refine rubrics—red text denotes inaccurate, automatically generated rubrics, while green text indicates expert-corrected versions; (3) **Model response**, where LLMs generate answers to hallucination-embedded inputs; and (4) **Evaluation**, where an LLM-as-a-judge framework assesses responses against the curated rubrics to determine hallucination rates.

over 160,000 patient summaries provides a rich and diverse foundation of authentic clinical content, ideal for testing an LLM's reasoning and information extraction capabilities. Our data generation process involved four key stages:

(1) Context-aware Adaptive Trap Injection. This forms the core of our adversarial data generation. We employ a powerful LLM generator, acting as a "misinformation architect", to perform the generation. For the MHB-Dialogue dataset, the generator is instructed to strategically embed "hallucination traps" by adaptively selecting one or more trap types from our hallucination traps (detailed in Table 1). This process is context-aware; the choice of trap is tailored to the dialogue's specific medical context to maximize plausibility. For instance, a discussion about hypertension might be subtly altered to include a reference to an outdated treatment guideline (Trap E1), whereas a conversation about a new symptom might introduce a fabricated diagnostic test (Trap A1). For the MHB-Report dataset, our goal was to create adversarial questions based on the source reports. Instead of modifying the source content, we use the generator LLM to architect questions that introduce subtle *factual vacuums* and *reasoning pitfalls*. Given a patient summary, postoperative note, or examination report as context, as well as the hallucination traps in Table 1, the generator was prompted to design a clinically plausible question that, while superficially relevant, could not be answered from the provided text.

(2) Progressive Induction Strategy. For the MHB-Dialogue dataset, instead of inserting a single, isolated false statement, the generator subtly restructures the semantics and adjusts the logic of the dialogue from both the user and the assistant. This weaves a web of suggestive context and misleading logical cues, creating a strong cognitive bias. The process culminates in a carefully crafted final question from the user, which acts as bait, designed to provoke the target model into generating a hallucinated response based on the previously embedded traps. For the MHB-Report dataset, the questions were engineered to tempt the target model into making unsupported inferences via progressive question structures. For example, a question might inquire about the results of a specific diagnostic test that was never mentioned (Trap A1), ask for a treatment recommendation based on a fabricated postoperative complication (Trap D1), or prompt a prognostic assessment that requires extrapolating far be-

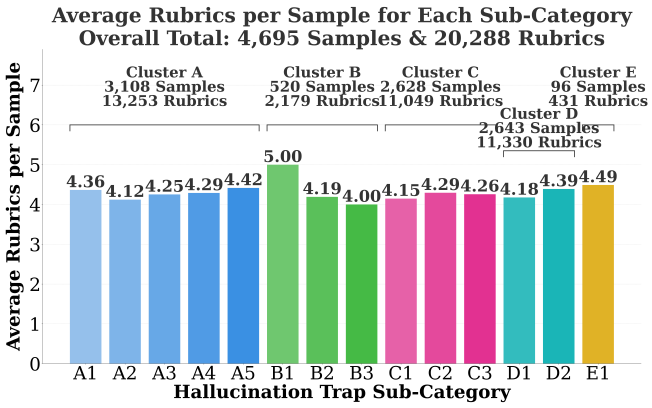


Figure 2: **Overview of the MHB Dataset.** The chart displays the distribution of rubrics and samples across hallucination traps. The benchmark comprises 4,695 samples and 20,288 rubrics, with each sample annotated by an average of four rubrics. The three largest clusters—A (3,108 samples), C (2,628 samples), and D (2,643 samples)—capture the most prevalent and clinically critical types of medical hallucinations. Additionally, the benchmark incorporates rarer yet clinically significant clusters, such as B (520 samples) and E (96 samples), to ensure a comprehensive evaluation of model robustness across the full spectrum of potential errors.

yond the available evidence (Trap C2).

(3) Automated Rubric Curation. In parallel with generating the trap-embedded dialogues and questions, the generator LLM also produces a structured JSON object containing a detailed evaluation rubric. This rubric explicitly defines the nature of the implanted hallucination and the precise criteria for a correct, non-hallucinatory response (e.g., "The model must identify 'Neurostatin' as a fabricated drug" or "The model should state that the report does not contain information on a PET scan"). This parallel generation of trap-embedded dialogue/questions, as well as their corresponding evaluation standards, creates a tightly coupled, closed-loop system, ensuring high internal consistency between the trap-embedded dialogue/questions and their traps, thus facilitating automated, scalable evaluation.

(4) Expert-led Quality Assurance. To ensure the clinical fidelity and rigorousness of our dataset, *we have instituted a multi-stage quality assurance protocol led by a panel of 60 professional, licensed physicians. These experts are recruited from the most prestigious Grade-A tertiary hospitals in China and possess extensive clinical experience across various specialties, including internal medicine, surgery, oncology, and pediatrics.* Our quality assurance involves a dual-annotator and cross-verification workflow. The initial review is conducted by two physicians independently. The **Primary Reviewer** performs the initial comprehensive assessment of the sample against a set of stringent criteria. Subsequently, a **Secondary Reviewer** independently verifies the primary reviewer's annotations and conclusions. Each sample is meticulously scrutinized against the following four core criteria:

- **Naturalness and Plausibility:** This criterion evaluates the subtlety and realism of the embedded trap. The expert reviewers assess whether the trap-infused dialogue or question flows naturally within a clinical context. A trap that is too obvious, artificial, or out-of-context would not constitute a fair or challenging test for an LLM.
- **Factual Correctness of the Trap:** Reviewers perform rigorous fact-checking to confirm that the "hallucinated" information is definitively and verifiably false. This process involves cross-referencing the information against authoritative medical sources, such as national clinical practice guidelines.
- **Logical Coherence:** The trap must be logically and semantically consistent with the surrounding text in the dialogue or medical report. A logically dissonant or self-contradictory trap can be easily detected through simple pattern recognition, failing to test a model's deeper reasoning capabilities.
- **Rubric-Trap Alignment:** This is a crucial check for evaluation integrity. The reviewer confirms that the model-generated evaluation rubric is precisely and unambiguously aligned with the specific hallucination trap embedded in the sample. A misaligned or vague rubric would render the automated evaluation process unreliable. The rubric must provide a clear standard for judging whether a model has successfully identified, understood, and corrected the specific error.

The quality assurance process is iterative. If a sample's generated rubric is deemed inadequate by either reviewer, it is rejected and regenerated. The new version then re-enters the review cycle from the beginning.

Hallucination Evaluation Method

To enable a scalable, consistent, and fine-grained assessment of LLM performance on our benchmark, we designed a robust evaluation methodology centered on an *LLM-as-a-judge* framework. This approach leverages a powerful judge model to automate the scoring process, guided by the expert-curated rubrics that were co-generated with each data sample. The entire evaluation process, from data construction to final judgment, is illustrated in Figure 1.

Automated Evaluation Pipeline. Our evaluation pipeline is executed for every sample in both the MHB-Dialogue and MHB-Report datasets. The process unfolds as follows:

(1) Response Generation. The target LLM is presented with the adversarial input. For MHB-Dialogue, this is the multi-turn conversation ending with a bait question. For MHB-Report, this is the combination of a clinical report and a hallucination-inducing question. The model's generated completion or answer is then collected.

(2) Judgment Assembly. A prompt is constructed for the judge LLM. This prompt bundles the complete context (the dialogue or report), the stimulus (the final question), the target model's response, and the specific, expert-validated evaluation rubric associated with that sample.

(3) Rubric-guided Judgment. The judge LLM is instructed to act as an impartial evaluator. Its sole task is to determine if the target model's response satisfies the criteria outlined in the provided rubric. The judge outputs its verdict in a structured JSON format containing two key fields:

- `criteria_met`: A boolean value indicating whether the response successfully avoids the hallucination and meets all rubric requirements. This serves as the primary quantitative metric for our analysis.
- `explanation`: A string containing a concise rationale for the judgment, referencing the rubric and explaining how the target model's response succeeded or failed. This provides qualitative insight into the model's behavior.

This automated pipeline establishes a closed-loop system where the criteria for generating a trap are the same as those for evaluating it, ensuring high fidelity and consistency. The structured output allows us to efficiently calculate hallucination rates across different models, hallucination traps (as defined in Table 1), and task formats.

Experiment

Experimental Settings

Dataset Construction Setting. The trigger-embedding process in both datasets was performed programmatically using the Google `Gemini 2.5-Pro` model, accessed via its official API and configured with default hyperparameters. *A detailed exposition of the prompt engineering strategies is provided in the supplementary material.*

Evaluation Setting. The `DeepSeek-R1-0528` model was employed as the "Judge" to assess the presence of hallucinations in the outputs of the evaluated models. To ensure the stability and reproducibility of the evaluation, the generative temperature of the "Judge" was set to a deterministic value of 0.5. This low-temperature setting minimizes output stochasticity, thereby yielding more consistent and reliable evaluation judgments across all assessed models.

Models Under Evaluation. To comprehensively assess the prevalence and nature of hallucinatory responses, we evaluated a suite of state-of-the-art (SOTA) large language models. This suite included several prominent proprietary models: `Gemini 2.5-Pro` (denoted as `Gemini`), `Claude-4-Sonnet` (denoted as `Claude`), `GPT-4.1-2025-4-14` (denoted as `GPT`). Additionally, we evaluated leading open-source models, specifically `Qwen3-235B-A22B` (denoted as `Qwen`), `DeepSeek-R1-0528` (denoted as `DeepSeek`), and `Llama-4-Scout-17B-16E-Instruct` (denoted as `Llama`). To ensure standardized and reproducible evaluation conditions, all models were accessed and queried via their official APIs.

Metric. The hallucination rate is calculated as the proportion of rubrics that are not passed out of the total rubrics assessed.

Experimental Results

This section presents the empirical findings from our evaluation of several LLMs on the MHB benchmark. The analysis quantifies model-specific hallucination rates across two sub-datasets. We report the hallucination rates.

Performance on MHB-Dialogue Dataset. In the MHB-Dialogue setting, we evaluated the propensity of a subset of models to hallucinate in a patient-clinician conversational interaction.

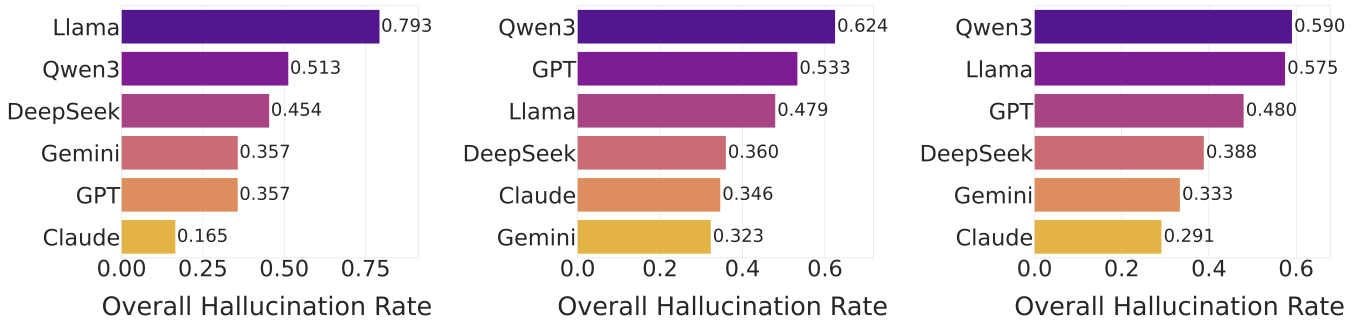
The overall hallucination rates are presented in Figure 3a. `Claude` demonstrated the highest factual robustness, exhibiting the lowest hallucination rate of 0.165. Conversely, `Llama` was the most susceptible to hallucination, with a rate of 0.793. Other models occupied a performance middleground, with `Qwen` at 0.513, `DeepSeek` at 0.454, and both `Gemini` and `GPT` recording identical rates of 0.357.

To dissect these aggregate findings, we performed a granular analysis of model performance across predefined trap categories in Figure 4a. The results revealed that implant trap types A1 (Non-existent Entity) and D2 (Source Fabrication) consistently induced a high frequency of hallucinations across most models, indicating these represent common vulnerabilities. `Llama` showed extreme susceptibility across nearly all implant types, with rates approaching or exceeding 0.8. In stark contrast, `Claude` maintained a low hallucination profile across all implant categories. The corresponding breakdown by hallucination trap cluster (see supplementary) further highlighted these disparities. Trap cluster A (Factual Discrepancy Traps) and D (Procedural Error Traps), in particular, proved exceptionally challenging for `Llama`, while `Claude` again demonstrated superior and consistent performance.

Performance on the MHB-Report Dataset. Our evaluation on the MHB-Report dataset, which benchmarks the task of processing and summarizing clinical reports, yielded a different performance hierarchy (Figure 3b). In this context, `Gemini` achieved the lowest hallucination rate at 0.323, closely followed by `Claude` (0.346) and `DeepSeek` (0.360). At the opposite end of the performance spectrum, `Qwen` registered the highest rate of 0.624, with `GPT` also showing significant vulnerability (0.533).

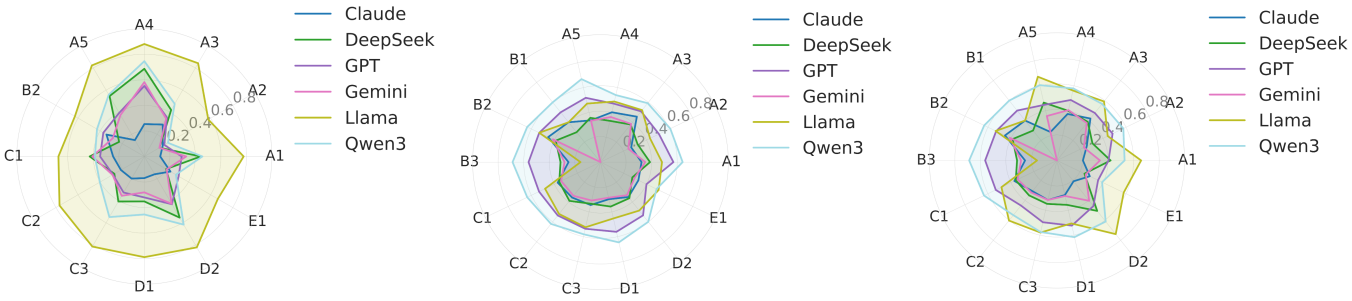
A dimensional analysis of these results on hallucination traps, visualized in Figure 4b, identified specific model failure modes. The implant trap type A3 (Incorrect Factual Claim) consistently triggered high hallucination rates across multiple models, most notably in `Qwen` and `GPT`. When analyzed by hallucination cluster (see supplementary), cluster E (Temporal Obsoletion Traps) emerged as a prominent vulnerability, particularly for `GPT` and `Llama`. Consistent with its strong overall performance on this dataset, `Gemini` maintained a robustly low hallucination profile across all evaluated implant types and hallucination clusters.

Aggregated Performance on the Combined MHB Dataset. To provide a comprehensive performance overview, we aggregated the results from both the dialogue and report sub-datasets. This combined analysis offers a holistic measure of model reliability across varied medical contexts. Figure 3c presents the overall hallucination rates on the complete MHB benchmark. `Claude` emerged as the most robust model overall, with the lowest aggregate hallucination rate of 0.291. It was followed by `Gemini` (0.333) and `DeepSeek` (0.388). Conversely, `Qwen` and



(a) MHB-Dialogue (b) MHB-Report (c) MHB-Combined

Figure 3: **Overall hallucination rates on the MHB dataset.** The charts show the aggregate hallucination rate for each model on the (a) dialogue, (b) report, and (c) combined subsets. Lower values indicate better performance.



(a) MHB-Dialogue (b) MHB-Report (c) MHB-Combined

Figure 4: **Dimensional analysis of hallucination rates by trap types.** Radar charts display hallucination rates (from 0 to 1.0) on the (a) MHB-Dialogue, (b) MHB-Report, and (c) combined datasets. For each category, points closer to the center indicate lower hallucination rates and therefore better performance.

`Llama` recorded the highest overall rates at 0.590 and 0.575, respectively.

The combined categorical analysis in Figure 4c reinforces the systemic nature of certain vulnerabilities. The radar plot for implant trap types confirms that categories such as D2 and C3 represent persistent challenges for the current generation of LLMs. Similarly, the analysis by hallucination cluster (see supplementary) identifies clusters A and D as areas of common weakness. The performance disparity between models was not uniform; while most models struggled with cluster A, C, and D, the magnitude of the hallucination rate was substantially larger for models like `Llama` and `Qwen` compared to the more resilient `Claude` and `Gemini`. This suggests that the identified vulnerabilities may stem from fundamental differences in model architecture, training data, or fine-tuning procedures.

Limitations

Our study has limitations. First, our binary metric only identifies the presence of hallucinations, not their clinical severity or potential harm; future work should prioritize risk-stratified metrics considering clinical impact. Second, we tested specific adversarial manipulations, and models may have other vulnerabilities. Finally, while model rankings will change as technology evolves, we contend that the systemic vulnerabilities and evaluation framework identified will remain relevant for guiding future development and safety testing.

Conclusion

Our evaluation shows contemporary LLMs are unreliable for autonomous, high-stakes clinical deployment. High hallucination rates, particularly from adversarial triggers, underscore the risks of unmonitored use. Consequently, clinical integration must mandate a "human-in-the-loop" approach, requiring expert verification of all outputs to ensure patient safety. This study provides a robust benchmark to guide this process and future development. We urge the research community to mitigate the high-risk failure modes identified herein to responsibly advance this technology in medicine.

Acknowledgements

This work was supported by the Key R&D Program of Zhejiang Province (No. 2025C01084) and the Ant Group Research Intern Program.

References

Agarwal, V.; Jin, Y.; Chandra, M.; Choudhury, M. D.; Kumar, S.; and Sastry, N. 2024. MedHalu: Hallucinations in

Responses to Healthcare Queries by Large Language Models. .

Al-Hakami, A. A.; Alsulaiman, A.; Al-Khalifah, A. S.; Al-Salman, N.; Alruwaili, M.; Alshammari, B.; and Al-Abdullah, A. 2024. Large Language Models in Healthcare and Medical Domain: A Review. *arXiv preprint arXiv:2401.06775*.

Arias-Duart, A.; Martin-Torres, P. A.; Hinjos, D.; Bernabeu-Perez, P.; Ganzabal, L. U.; Gonzalez Mallo, M.; Gururajan, A. K.; Lopez-Cuena, E.; Alvarez-Napagao, S.; and Garcia-Gasulla, D. 2025. Automatic Evaluation of Healthcare LLMs Beyond Question-Answering. *arXiv preprint arXiv:2502.06666*.

Arora, R. K.; Wei, J.; Hicks, R. S.; Bowman, P.; Quiñonero-Candela, J.; Tsimpourlas, F.; Sharman, M.; Shah, M.; Vallone, A.; Beutel, A.; Heidecke, J.; and Singhal, K. 2025. HealthBench: Evaluating Large Language Models Towards Improved Human Health. .

Bang, Y.; Ji, Z.; Schelten, A.; Hartshorn, A.; Fowler, T.; Zhang, C.; Cancedda, N.; and Fung, P. 2025. HalluLens: LLM Hallucination Benchmark. .

Chang, A.; Huang, L.; Bhatia, P.; Kass-Hout, T.; Ma, F.; and Xiao, C. 2025. MedHEval: Benchmarking Hallucinations and Mitigation Strategies in Medical Large Vision-Language Models. *arXiv preprint arXiv:2503.02157*.

Chen, J.; Yang, D.; Wu, T.; Jiang, Y.; Hou, X.; Li, M.; Wang, S.; Xiao, D.; Li, K.; and Zhang, L. 2024. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.

Chen, Q.; Hu, Y.; Peng, X.; Xie, Q.; Jin, Q.; Gilson, A.; Singer, M. B.; Ai, X.; Lai, P.-T.; Wang, Z.; et al. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature communications*, 16(1): 3280.

Das, A. B.; Ahmed, S.; and Sakib, S. K. 2025. Hallucinations and key information extraction in medical texts: A comprehensive assessment of open-source large language models. *arXiv preprint arXiv:2504.19061*.

Dou, C.; Zhang, Y.; Chen, Y.; Jin, Z.; Jiao, W.; Zhao, H.; and Huang, Y. 2024. Detection, diagnosis, and explanation: A benchmark for chinese medial hallucination evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 4784–4794.

Fleming, S. L.; Lozano, A.; Haberkorn, W. J.; Jindal, J. A.; Reis, E.; Thapa, R.; Blankemeier, L.; Genkins, J. Z.; Steinberg, E.; Nayak, A.; et al. 2024. Medalign: A clinician-generated dataset for instruction following with electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22021–22030.

Han, T.; Kumar, A.; Agarwal, C.; and Lakkaraju, H. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Advances in Neural Information Processing Systems*, 37: 33423–33454.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2020. What Disease Does This Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. .

Kim, Y.; Jeong, H.; Chen, S.; Li, S. S.; Lu, M.; Alhamoud, K.; Mun, J.; Grau, C.; Jung, M.; Gameiro, R.; et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Li, J.; Cheng, X.; Zhao, X.; Nie, J.-Y.; and Wen, J.-R. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6449–6464. Singapore: Association for Computational Linguistics.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252. Association for Computational Linguistics.

Liu, M.; Ding, J.; Xu, J.; Hu, W.; Li, X.; Zhu, L.; Bai, Z.; Shi, X.; Wang, B.; Song, H.; Liu, P.; Zhang, X.; Wang, S.; Li, K.; Wang, H.; Ruan, T.; Huang, X.; Sun, X.; and Zhang, S. 2024. MedBench: A Comprehensive, Standardized, and Reliable Benchmarking System for Evaluating Chinese Medical Large Language Models. .

Nguyen, D.; Ho, M. K.; Ta, H.; Nguyen, T. T.; Chen, Q.; Rav, K.; Dang, Q. D.; Ramchandre, S.; Phung, S. L.; Liao, Z.; et al. 2025. Localizing Before Answering: A Hallucination Evaluation Benchmark for Grounded Medical Multimodal LLMs. *arXiv preprint arXiv:2505.00744*.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2023. Medhalt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Xu, Y.; Cai, T.; Jiang, J.; and Song, X. 2024. Face4rag: Factual consistency evaluation for retrieval augmented generation in chinese. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 6083–6094.

Yan, Q.; Yuan, Y.; Hu, X.; Wang, Y.; Xu, J.; Li, J.; Fu, C.-W.; and Heng, P.-A. 2025. MedHallTune: An Instruction-Tuning Benchmark for Mitigating Medical Hallucination in Vision-Language Models. *arXiv preprint arXiv:2502.20780*.

Zhao, Z.; Jin, Q.; Chen, F.; Peng, T.; and Yu, S. 2023. A large-scale dataset of patient summaries for retrieval-based clinical decision support systems. *Scientific Data*, 10(1): 909.

Zuo, K.; and Jiang, Y. 2024. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *arXiv preprint arXiv:2412.18947*.