Can We Trust AI Doctors? A Survey of Medical Hallucination in Large Language and Large Vision-Language Models

Zhihong Zhu[1], **Yunyan Zhang**[1], **Xianwei Zhuang**[2], **Fan Zhang**[3], **Zhongwei Wan**[4], **Yuyan Chen**[5], **Qingqing Long**[2], **Yefeng Zheng**[6], **Xian Wu**[1,†]
[1]Tencent Jarvis Lab [2]Peking University [3]The Chinese University of Hong Kong
[4]The Ohio State University [5]Fudan University [6]Westlake University
zhihongzhu@stu.pku.edu.cn kevinxwu@tencent.com

Abstract

Hallucination has emerged as a critical challenge for large language models (LLMs) and large vision-language models (LVLMs), particularly in high-stakes medical applications. Despite its significance, dedicated research on medical hallucination remains unexplored. In this survey, we first provide a unified perspective on medical hallucination for both LLMs and LVLMs, and delve into its causes. Subsequently, we review recent advancements in detecting, evaluating, and mitigating medical hallucinations, offering a comprehensive overview of evaluation benchmarks, metrics, and strategies developed to tackle this issue. Moreover, we delineate the current challenges and delve into new frontiers, thereby shedding light on future research. We hope this work can provide the community with quick access and spur breakthrough research in this area.

1 Introduction

In the rapidly evolving realm of artificial intelligence, large language models (LLMs) and large vision-language models (LVLMs) have achieved significant success and demonstrated promising capabilities across a wide range of applications (OpenAI, 2022; OpenAI et al., 2023; Li et al., 2023a; Zhang et al., 2025a). Notably, the integration of these large foundational models into real-world medical practice holds immense potential (Singhal et al., 2023; Wu et al., 2024c; Wan et al., 2023; Zhang et al., 2025b), as they can alleviate doctors' workloads, reduce costs, and increase clinical accessibility. Medical-specialized LLMs such as Med-PaLM 2 (Singhal et al., 2025), Med-Gemini (Saab et al., 2024) and MedGemma[1] have already demonstrated impressive performance on a variety of medical benchmarks.

†Corresponding author.

[1]https://github.com/google-health/medgemma

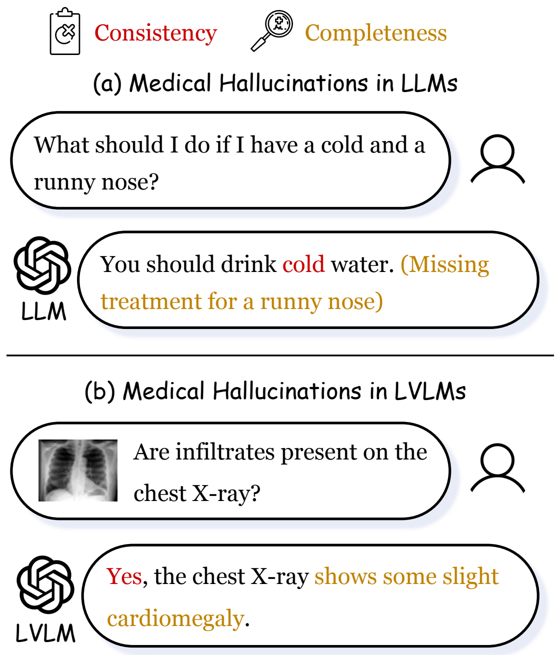

Figure 1: Two cases of medical hallucinations in LLMs (a) and LVLMs (b), where the response in yellow indicates *completeness* and in red denotes *consistency* hallucination. (a) The model erroneously advises drinking cold water instead of medically recommended warm fluids, while failing to address the presented symptom of runny nose. (b) The model falsely identifies pulmonary infiltrates despite their absence, with a comment on cardiomegaly unrelated to clinical inquiry.

General vs. Medical. However, a critical barrier to the clinical deployment of these models is their *hallucination* propensity. In general LLMs, hallucination typically refers to content that is nonsensical or unfaithful to the source material, categorized as *factuality* or *faithfulness* issues (Ji et al., 2023a; Huang et al., 2023a). For general LVLMs, it manifests as mismatches between imaging evidence and textual outputs, classified into *object*, *attribute*, and *relation* subtypes (Liu et al., 2024a; Bai et al., 2024). Unfortunately, these existing taxonomies present two limitations for medical applications: *(1)* incomplete coverage due to non-

Findings of the Association for Computational Linguistics: ACL 2025, pages 6748–6769
July 27 - August 1, 2025

Figure 2: The primary causes of hallucination within the medical domain.

overlapping hallucination types between LLMs and LVLMs, and *(2)* domain misalignment where general-domain frameworks fail to capture critical clinical risks in medical contexts – *missed diagnoses* and *misdiagnoses* (Schiff et al., 2009).

Definition. Following these premises, we first introduce a unified taxonomy that bridges LLM and LVLM hallucinations through (*cf.* Figure 1): *(1) Consistency* corresponds to *misdiagnoses*: Ensuring generated content aligns with medical evidence (image/text), clinical context, and domain knowledge; *(2) Completeness* corresponds to *missed diagnoses*: Encompassing both omission risks (missing critical findings) and extraneous content (irrelevant information that distorts clinical focus).

Motivation. Unlike general applications where hallucinatory contents may be somewhat tolerable, hallucination in high-stakes medical contexts can lead to severe consequences, such as misdiagnoses and inappropriate treatments (Vishwanath et al., 2024; Chen et al., 2024b). However, there is a lack of a systematic overview and summary of hallucination under LLMs and LVLMs in the medical field, which hinders further progress in this area.

To bridge this gap, we conduct a comprehensive analysis of medical hallucination across both LLMs and LVLMs through a lifecycle-oriented (Mündler et al., 2023) framework encompassing four critical dimensions: *(1)* cause analysis (to identify underlying triggers), *(2)* detection methodologies (techniques for identifying hallucinatory content), *(3)* evaluation protocols (metrics and benchmarks for systematic assessment), and *(4)* mitigation strategies (interventions to reduce occurrences).

Contribution. *(1) Comprehensive Survey*: To the best of our knowledge, this is the first effort to introduce a comprehensive survey dedicated to medical hallucination. *(2) Meticulous Taxonomy*: We introduce a unified and meticulous taxonomy towards both LLMs and LVLMs (*cf.* Figure 2 and Figure 3); *(3) Challenges and Frontiers*: We discuss existing challenges and new frontiers in medical hallucination, and shed light on future research; *(4)*

Survey	Medical Tailored?	Discussion Focus	
		LLMs?	LVLMs?
Huang et al. (2023a)	✗	✔	✗
Zhang et al. (2023c)	✗	✔	✗
Bai et al. (2024)	✗	✗	✔
Liu et al. (2024a)	✗	✗	✔
Sahoo et al. (2024)	✗	✔	✔
Ours (Dec. 2024)	✔	✔	✔

Table 1: Comparison with existing related surveys.

Abundant Resources: We publicly release and continuously maintain relevant resources to facilitate future research in this field.[2]

Organization. In the following, we first present the hallucination-related surveys (§2). We then provide an introduction to the causes of medical hallucination (§3). Next, we discuss the detection, evaluation and mitigation of medical hallucination in LLMs (§4) and LVLMs (§5). Finally, we explore existing challenges (§6) and frontier research (§7).

2 Related Surveys

What did the previous review on hallucination discuss? Existing surveys on hallucination (Huang et al., 2023a; Zhang et al., 2023c; Bai et al., 2024; Liu et al., 2024a) (*cf.* Table 1) primarily address general-domain challenges. A series of recent studies (Sun et al., 2024; Jing and Du, 2024; Chen et al., 2024e; Jing et al., 2024; Liang et al., 2024) have made remarkable progress in mitigating hallucinations in LLMs or LVLMs. While Sahoo et al. (2024) extended the discussion to both LLMs and LVLMs, their analysis remains confined to generic detection and mitigation techniques. Crucially, the investigation of hallucinations in medical contexts remains absent. We also note two concurrent studies (Wang et al., 2025; Kim et al., 2025) that were completed after our work and share similar interests in medical hallucination.

Why a survey on medical hallucination? Beyond the low tolerance for errors, the need to

[2] `https://github.com/Zhihong-Zhu/Medical_Hallucination_Survey`

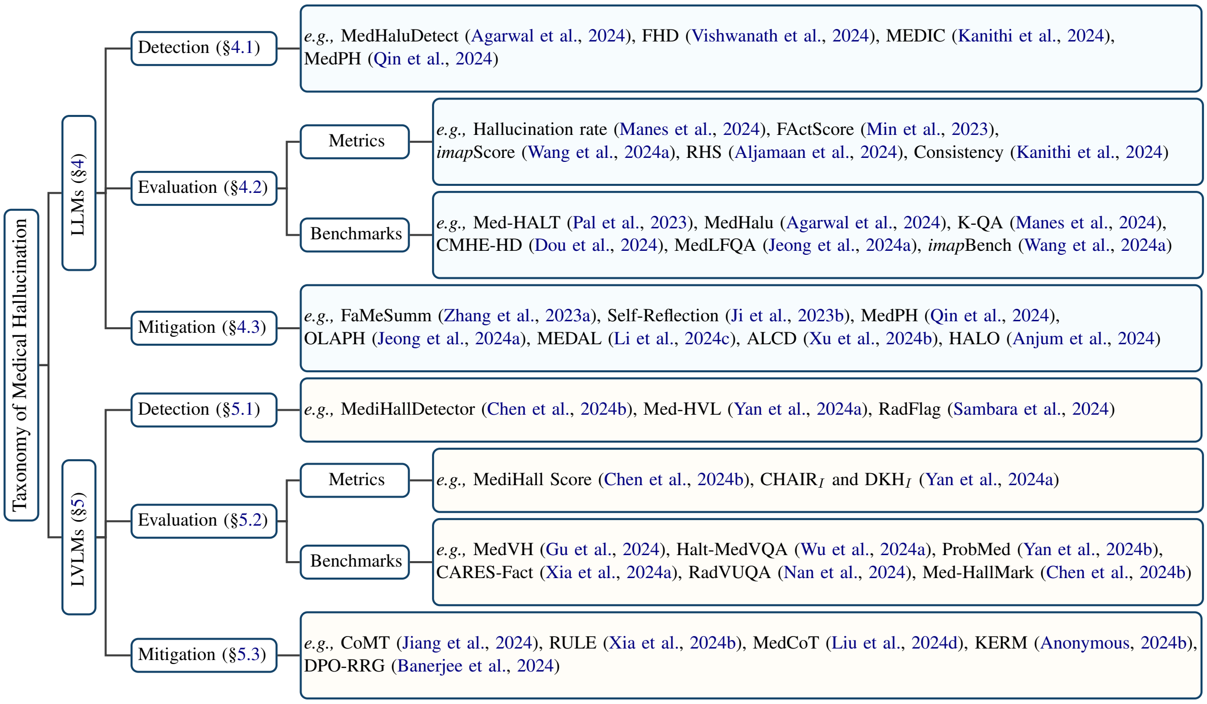

Figure 3: The taxonomy of medical hallucination in terms of detection, evaluation and mitigation. We primarily catalog recent research conducted in the era of LLMs and LVLMs.

study medical hallucinations is further emphasized by several factors compared to the general domain: *(1)* Medical data introduces unique hallucination triggers not found in general domains (Vishwanath et al., 2024). *(2)* Clinical deployment constraints (Hager et al., 2024) require tailored evaluation protocols that go beyond conventional metrics. *(3)* Specialized mitigation strategies should be developed, particularly leveraging prior clinical domain knowledge (Van Veen et al., 2024).

In light of these, we present the first systematic examination of medical hallucinations across the complete lifecycle - from cause analysis through detection and evaluation to mitigation.

3 Causes of Medical Hallucination

In this section, we briefly discuss the causes of medical hallucination in both LLMs and LVLMs (*cf.* Figure 2), categorizing them into data level (§3.1), training level (§3.2) and inference level (§3.3).

3.1 Data Level

The strict privacy policies exacerbate insufficient and imbalanced training data (Jiang et al., 2024), which is one of the primary factors contributing to medical hallucination (Guo and Terzopoulos, 2024; Chen et al., 2024b). Many pathologies are underrepresented in medical datasets, making it challenging for models trained on large-scale medical data to effectively learn the features of these less common conditions (Liu et al., 2024a).

Other potential data issues include misconceptions (Lin et al., 2022), social biases (Ladhak et al., 2023), and duplication biases (Lee et al., 2022). Moreover, models may generate responses that are inconsistent with real-time information due to their reliance on static world knowledge acquired during pre-training (Onoe et al., 2022; Addlesee, 2024).

Remarks. *The issues on data quantity and biases are amplified in the medical domain, distinguishing it from general domains (Koetzier et al., 2024).*

3.2 Training Level

Due to the lag in feature extraction and architectural design compared to general domains (Yan et al., 2024b; Chen et al., 2024b), existing methods struggle to effectively handle fine-grained medical inquiries. Additionally, empirical studies (Pal et al., 2023) highlighted a potential detrimental effect on a model's ability to control hallucinations following Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). This issue can be attributed to belief misalignment, or sycophancy (Burns et al., 2022), where the model generates responses based on user satisfaction rather than factual accuracy, due to training on datasets prioritizing the former.

Besides, exposure bias (Ranzato et al., 2015) creates a divergence between the model's training and inference behavior (Yan et al., 2024b; Anjum et al., 2024). This discrepancy forces the model to operate beyond its learned knowledge boundaries, a phenomenon termed capability misalignment.

Remarks. *Factors such as the lag in tailored model design, the absence of domain-specific training, and inadequate post-training processes all contribute to the occurrence of medical hallucination.*

3.3 Inference Level

Recent studies have showcased that Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) can enhance medical factual accuracy. However, excessive reliance on external knowledge may lead to incorrect answers and hallucinations (Xia et al., 2024b; Ahmad et al., 2023). Moreover, decoders may attend to the wrong part of the encoded input, leading to erroneous generation (Ji et al., 2023a).

In contrast, an incorrect token generated through stochastic sampling (Chuang et al., 2023) can often trigger a chain of subsequent errors (Xiang et al., 2023), which can severely undermine applications such as medical summarization (Vishwanath et al., 2024). The prompts provided to models also play a crucial role in generating hallucinations. Poorly constructed prompts (Yu et al., 2022) may fail to fully capture the context of a query, resulting in insufficient context attention (Shi et al., 2023).

Remarks. *Tool use, decoding strategies, and prompt design are key research areas in addressing medical hallucinations during inference.*

4 Medical Hallucination in LLMs

This section provides a holistic review of medical hallucination in existing LLMs, including detection (§4.1), evaluation (§4.2) and mitigation (§4.3).

4.1 Detection

Current focus in LLMs has primarily been on detecting *consistency* hallucination through domain-adapted verification frameworks. MedHaluDetect (Agarwal et al., 2024) triangulated input-, context-, and fact-conflict hallucinations via collaborative LLM-expert-layperson assessments, explicitly modeling medical QA inconsistencies. Besides, MEDIC (Kanithi et al., 2024) synthesized adversarial yes/no questions from source documents through automated cross-examination to expose contradictions in generated summaries, which mitigates reliance on scarce clinical annotations.

The integration of structured medical knowledge further differentiates these methods from general-domain approaches. Thereinto, FHD (Vishwanath et al., 2024) combined rule-based event extraction (Hypercube) with LLM-driven semantic analysis (GPT-4) to identify inconsistencies in medical summaries. Furthermore, MedPH (Qin et al., 2024) addressed temporal inconsistencies in patient dialogues by constructing entity graphs with structural entropy, explicitly modeling discrepancies between reported symptoms and inferred diagnoses.

Remarks. *Medical hallucination detection in LLMs diverges from general-domain approaches by (1) employing expert-guided synthetic evaluation to address data scarcity, (2) grounding in clinical ontologies to resolve ambiguities, and (3) prioritizing critical clinical errors over generic inaccuracies. Unlike conventional methods relying on probabilistic uncertainty, medical adaptations necessitate hybrid architectures integrating biomedical knowledge graphs and risk-aware workflows.*

4.2 Evaluation

Metrics. Evaluating medical hallucinations in LLMs necessitates specialized metrics that address the high stakes of clinical errors and domain-specific knowledge granularity (Dou et al., 2024). While traditional NLP metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) measure surface-level text similarity, they fail to capture clinically critical inconsistencies. Recent efforts have shifted toward hallucination-centric evaluation protocols that prioritize factual accuracy over lexical overlap. For instance, Hallucination Rate (Manes et al., 2024) directly quantifies contradictions between model outputs and ground-truth statements.

Emerging medical-specific metrics further incorporate structured clinical knowledge to address ambiguities in terminology and causality. The Reference Hallucination Score (RHS) (Aljamaan et al., 2024) evaluates AI-generated medical references by auditing seven identifier categories. Similarly, *imap*Score (Wang et al., 2024a) validates the alignment of structured medical term-value pairs. These methods contrast with general-domain fact-checking tools like FActScore (Min et al., 2023) by grounding evaluations in clinical ontologies rather than open-world knowledge. Concurrently, to bypass annotation scarcity, MEDIC's consistency score (Kanithi et al., 2024) measures hallucination severity as the percentage of summary-derived questions unanswerable from source documents.

Benchmark	Data Size	Task Type	Evaluation Metric	Hallucination	
				Consistency	*Completeness*
Med-HALT (Pal et al., 2023)	59,254	Dis	Acc, Pointwise Score	✔	✔
MedHalu (Agarwal et al., 2024)	2,077	Dis	Acc, P, R, F1	✔	✗
CMHE-HD (Dou et al., 2024)	2,000	Dis	Acc	✔	✗
K-QA (Manes et al., 2024)	201	Gen	Comp, Hall	✔	✔
MedLFQA (Jeong et al., 2024a)	4,948	Gen	R-1/2/L, Comp, Hall, BERTScore, BLEU(RT)	✔	✔
*imap*Bench (Wang et al., 2024a)	5,001	Gen	*imap*Score, Human	✔	✔

Table 2: Medical hallucination benchmarks for LLMs. Dis: Discriminative; Gen: Generative; Acc: Accuracy; P: Precision; R: Recall; Comp: Comprehensiveness; Hall: Hallucination rate; R-1/2/L: ROUGE-1/2/L.

Benchmarks. Table 2 synthesizes representative medical hallucination benchmarks for LLMs, and we also present detailed baselines and their underlying data sources in Appendix Table 6.

1) Discriminative. Med-HALT (Pal et al., 2023) categorizes hallucinations into reasoning and memory-based types (Vilares and Gómez-Rodríguez, 2019; Jin et al., 2021; Pal et al., 2022), assessing consistency and completeness hallucination. MedHalu (Agarwal et al., 2024) focuses on hallucinations in real-world healthcare queries, addressing contextual and data diversity through expert-curated (Zhu et al., 2019) and consumer-oriented (Abacha et al., 2017, 2019) datasets. Additionally, CMHE-HD (Dou et al., 2024) assesses LLMs' ability to detect misinformation in doctor-patient dialogues, which includes hallucination-free samples from cMedQA2 (Zhang et al., 2018) and CMD, and hallucinated samples generated by Llama2 (Touvron et al., 2023b) and ChatGPT.

2) Generative. K-QA (Manes et al., 2024) evaluates LLMs on real-world medical QA from K Health[3], with two metrics: comprehensiveness to measure essential statement coverage, and hallucination rate to assess contradictions with the truth. Based on K-QA, MedLFQA (Jeong et al., 2024a) introduces "Must Have" and "Nice to Have" statements for automatic evaluation, to prevent hallucinations and ensure factual accuracy. *imap*Bench (Wang et al., 2024a) evaluates the factual correctness of LLM-generated QA responses, focusing on both consistency and completeness.

Remarks. *On one hand, the newly proposed metrics tailored to medical hallucination still primarily target consistency hallucination; on the other hand, to reduce reliance on medical exams, generative benchmarks have incorporated extensive annotation efforts involving both LLMs and human labor.*

[3] https://khealth.com/

4.3 Mitigation

Due to space limitations, the quantitative results and application can be found in Appendix Table 4.

Mitigation at the Data Level. Data-level innovations in medical LLMs address medical corpora limitations through targeted synthesis and selection. For instance, OLAPH (Jeong et al., 2024a) employed self-generated medical QA datasets (Chen et al., 2024h; Wu et al., 2024d) to bypass dependency on scarce annotated resources, coupled with metric-guided preference alignment that prioritizes clinically coherent responses. This contrasts with conventional data augmentation, by directly countering medical-specific hallucination drivers through context-aware synthetic examples.

Mitigation at the Training Level. Training-level mitigation solutions demonstrate a systematic shift toward medical knowledge integration and error containment. Thereinto, FaMeSumm (Zhang et al., 2023a) enhanced faithfulness in medical summarization by combining contrastive learning (Khosla et al., 2020; Cao and Wang, 2021) and medical knowledge incorporation. To mitigate hallucination in medical information extraction, Xu et al. (2024b) introduced ALCD, which decouples the identification and classification processes (Khot et al., 2023; Bian et al., 2023) during training.

Mitigation at the Inference Level. Based on the level of access required to the model, inference-level medical hallucination techniques can be classified into three main categories (Liu et al., 2024e; Huang et al., 2023b): *black-box* (using only generated outputs), *gray-box* (using output probabilities), and *white-box* (using internal components).

1) Black-box. Black-box methods prioritize clinical safety through iterative verification and knowledge grounding. Techniques like MEDAL (Li et al., 2024c) combined self-examination mechanisms like (Ji et al., 2023b) with

Benchmark	Data Size	Task Type	Evaluation Metric	Hallucination	
				Consistency	*Completeness*
MedVH (Gu et al., 2024)	-	Dis & Gen	Acc, CHAIR	✔	✗
Halt-MedVQA (Wu et al., 2024a)	2,359	Dis	Acc	✔	✗
Med-HallMark (Chen et al., 2024b)	7,341	Gen	BERTScore, R-1/2/L, BLEU, METEOR, MediHall Score, Acc	✔	✔
CARES-Fact (Xia et al., 2024a)	-	Dis	Acc	✔	✗
RadVUQA (Nan et al., 2024)	193,662	Dis & Gen	LLM-as-a-judge	✗	✔
ProbMed (Yan et al., 2024b)	57,132	Dis	Acc	✔	✗

Table 3: Medical hallucination benchmarks for LVLMs. Dis: Discriminative; Gen: Generative; Acc: Accuracy; R-1/2/L: ROUGE-1/2/L. '-' denotes missing details from the original publication.

synthetic non-factual summaries (Donahue et al., 2020) during post-processing, directly countering the limited availability of error-corrected medical corpora. Similarly, retrieval-augmented approaches such as HALO (Anjum et al., 2024) addressed medical specificity by dynamically expanding query perspectives to mitigate diagnostic tunnel vision.

2) Gray-box. Gray-box strategies reveal a distinctive focus on preventing cascading errors in clinical reasoning chains. For instance, ALCD (Xu et al., 2024b) adopted contrastive decoding through task-specific token masking that mimics clinical workflows where differential diagnosis precedes final classification. Combined with adaptive constraint strategy (Li et al., 2023b; Chuang et al., 2024), ALCD effectively adjusts the scale and scope of contrastive tokens, while minimizing the impact on other inherent abilities in LLMs.

3) White-box. White-box approaches uniquely operationalize clinical communication principles through architectural interventions. For example, MedPH (Qin et al., 2024) introduced proactive clarification generation (Rao, 2017) when detecting uncertain entity responses, mirroring clinicians' verification protocols during patient interviews.

Remarks. *Significant efforts have been made in mitigating medical hallucinations in LLMs, including the use of synthetic data at the data level, modifications and enrichment of training objectives at the training level, and self-correction, RAG, and contrastive decoding at the inference level.*

5 Medical Hallucination in LVLMs

This section delves into medical hallucination in existing LVLMs, highlighting their similarities and differences with hallucination in LLMs.

5.1 Detection

Recent advances in detecting medical hallucinations in LVLMs reveal two predominant strategies that address the unique challenges of clinical multimodal alignment. The first approach focuses on content verification through domain-specific cross-modal grounding, exemplified by MediHallDetector (Chen et al., 2024b) and Med-HVL (Yan et al., 2024a). In detail, MediHallDetector tripartited the classification of medical images, prompts, and answers, while Med-HVL directly contrasted extracted clinical entities against ground truth annotations. The second strategy emphasizes confidence estimation in clinical assertions, as demonstrated by RadFlag (Sambara et al., 2024) which identified inconsistencies across sampled report generations.

Remarks. *Unlike general-domain hallucination detection that can rely on commonsense validation, medical approaches must integrate domain knowledge to distinguish between clinically plausible inferences and factual errors. The progression from basic inconsistency detection (RadFlag) to sophisticated clinical concept verification (Med-HVL) reflects a paradigm shift toward expert-informed hallucination identification in medical LVLMs.*

5.2 Evaluation

Metrics. Existing evaluation metrics of LVLMs such as CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023c) focus on "object hallucination" in general domains, but fail to capture the complexities of medical hallucination.

To overcome this, MediHall Score (Chen et al., 2024b) was introduced as a fine-grained metric for tasks like Medical VQA and Imaging Report Generation (IRG), categorizing hallucinations by severity and aggregating scores at the sentence or answer level. Building on CHAIR, Yan et al. (2024a) refined object hallucination detection by calculating the proportion of objects mentioned in captions but not present in the image, and introduced DKH_I to evaluate domain knowledge hallucinations, particularly those arising from erroneous diagnoses due to biases inherited from the pre-trained LLM.

Benchmarks. Table 3 summarizes representative medical hallucination benchmarks for LVLMs, and we also present detailed baselines and their underlying data sources in Appendix Table 7.

1) Discriminative. Wu et al. (2024a) introduced Halt-MedVQA, a benchmark that modifies existing VQA datasets to include scenarios like fake questions, "None of the Above" answers, and image swaps (He et al., 2020a), to assess models' robustness in handling hallucinations. Yan et al. (2024b) introduced ProbMed for Med-VQA, emphasizing diagnostic capabilities under adversarial conditions in *consistency* hallucination. Additionally, Xia et al. (2024a) introduced the CARES-Fact benchmark, encompassing a broad spectrum of medical image modalities and anatomical regions (Johnson et al., 2019b; Demner-Fushman et al., 2016; Luo et al., 2024b; Lin et al., 2023).

2) Generative. Chen et al. (2024b) introduced the first multi-modal healthcare benchmark (Johnson et al., 2019a; Wang et al., 2017) Med-HallMark for hallucination detection including open-ended medical VQA (Liu et al., 2021; Lau et al., 2018) and image report generation (Johnson et al., 2019a; Wang et al., 2017). Med-HallMark offers hallucination data, including ground truth, LVLM outputs, and detailed annotations of hallucination types.

3) Discriminative & Generative. MedVH (Gu et al., 2024) evaluates multi-choice VQA and resistance to hallucinations in long context responses, which is constructed by several chest X-ray datasets (Zhang et al., 2023b; He et al., 2020b; Ben Abacha et al., 2021; Hu et al., 2023). Furthermore, Nan et al. (2024) presented Rad-VUQA, a benchmark incorporating both CT and MR datasets (Wasserthal et al., 2023; D'Antonoli et al., 2024; Xing et al., 2023; Soares et al., 2020).

Remarks. *Regarding metrics, the newly proposed metrics are derived from "object hallucination" in the general domain but lack deeper exploration such as "relation hallucination" (Wu et al., 2024b) in the medical domain. Regarding benchmarks, they predominantly focus on evaluating consistency hallucination, with comparatively less effort devoted to addressing completeness hallucination.*

5.3 Mitigation

Due to space limitations, the quantitative results and application can be found in Appendix Table 5.

Mitigation at the Data Level. Mitigation approaches in LVLMs increasingly focus on structural medical knowledge integration rather than generic visual-language pairing. CoMT (Jiang et al., 2024) hierarchically decomposed radiological reasoning into chain-of-thought QA pairs, mirroring radiologists' diagnostic workflows. Similarly, knowledge-enhanced retrieval systems like KERM (Anonymous, 2024b) mitigate modality-specific hallucinations through dynamic medical fact retrieval, prioritizing anatomical-pathological correlations over generic visual concepts.

Mitigation at the Training Level. Training-level mitigation methods in LVLMs shift toward medical error containment through constrained optimization. To address over-reliance in retrieval-augmented medical LVLMs, Xia et al. (2024b) proposed knowledge-balanced preference tuning to mitigate over-reliance on retrieved contexts. Parallel advancements in reward modeling, such as dual-level assessment frameworks (Anonymous, 2024b), enforce not just label accuracy but clinical narrative coherence in radiology reporting. The medical-adapted DPO methods (Banerjee et al., 2024) suppresses hallucinations by penalizing both textual deviations and visual-clinical incongruities, demonstrating how medical LVLM mitigation inherits yet expands upon LLM techniques through modality-specific constraints.

Mitigation at the Inference Level. At the inference level, medical hallucination mitigation approaches in LVLMs predominantly employ black-box strategies to improve model generalizationality. Thereinto, CoMT (Jiang et al., 2024) decomposes unstructured medical reports into fine-grained cues, which are then organized into a coherent chain of diagnostic reasoning. CoMT facilitates inductive reasoning based on detailed local features, thereby mitigating hallucinations and enhancing the reliability of the generated reports. MedCoT (Liu et al., 2024d) improves medical VQA accuracy and reliability through a three-stage inference pipeline. A consensus is reached through voting among trained Mixture of Experts (MoE) (Zhou et al., 2022; Cai et al., 2024), yielding the final diagnosis.

Remarks. *Chain-of-thought approaches (Wei et al., 2022) are frequently employed in LVLMs to mitigate medical hallucination, including fine-grained QA construction at the data level and multimodal reasoning during the inference phase. Additionally, variants of DPO and RLHF have been explored during the training phase to further enhance model accuracy and reduce hallucinations.*

6 Challenges

This section introduces the existing challenges of LLMs and LVLMs for medical hallucination in terms of data, model, and evaluation perspectives.

Data Control. Existing methods rarely address the impact of data bias in mitigating medical hallucination (Pham and Vo, 2024; Hegselmann et al., 2024). Evidently, various data biases can lead LLMs and LVLMs to learn idiosyncratic patterns tied to specific modalities and labels (Han et al., 2023b; Bhardwaj et al., 2023), rather than capturing the holistic semantics of instance. Thus, it is important to address hallucinations caused by the frequency or distribution of instances in the data (Cui et al., 2023; Zhu et al., 2024).

Due to the scarcity of medical data, synthetic data is extensively utilized in the medical domain (Koetzier et al., 2024; Mishra et al., 2023). However, such data is prone to generating scenarios that are inconsistent with real-world reality (Burgess et al., 2024). Consequently, it is essential to develop more effective data synthesis and filtering algorithms (Xie et al., 2024; Zhang et al., 2024a; Luo et al., 2024a) to enhance both the quantity and quality of data for medical hallucination.

Model Design. Despite promising results achieved, existing training-level hallucination mitigation methods remain focused on coarse-grained alignment and instance-level instruction tuning (Wang et al., 2024b). This focus limits their ability to comprehend complex symptoms and intricate details that are essential for models to suppress undesired outputs. Developing more meticulous supervision objectives (Anonymous, 2024a; Chen et al., 2023a) holds significant potential for medical hallucination mitigation.

With the advancement of more sophisticated LLMs and LVLMs, mitigating medical hallucination can be enhanced through the framework design (Wang et al., 2024c). Typical frameworks include multi-agent debate (Du et al., 2024), voting system (Wang et al., 2023), multi-disciplinary collaboration (Tang et al., 2024), group discussion (Chen et al., 2023c), and negotiation mechanism (Fu et al., 2023). Preliminary attempts have been made in areas such as medical decision-making (Kim et al., 2024; Li et al., 2024b; Ke et al., 2024). Developing more reliable frameworks remains a crucial direction for future research.

Evaluation Protocol. Tables 6 and 7 respectively summarize the underlying data sources of existing medical hallucination benchmarks. It is evident that current benchmarks primarily focus on healthcare queries and CT imaging reports (Royer et al., 2024; Tu et al., 2024). To comprehensively address medical hallucination, a broader range of anatomical structures should be incorporated including the brain, eyes, heart, and chest, among others.

Furthermore, the evaluation of medical hallucinations should extend beyond purely textual and text-image modalities, as hallucinations frequently occur in other modalities, such as medical video QA (Saab et al., 2024), surgery video understanding (Bai et al., 2023c), and streamlining medical transcription(Ebadi et al., 2024). Therefore, future efforts should incorporate video and audio modalities to comprehensively evaluate medical hallucination (Ma et al., 2024; Sun et al., 2025).

7 Frontiers

Despite promising advancements, research on medical hallucination remains in its early stages, leaving ample room for further development. In the following, we outline these areas for future research.

Efficiency. Inference-level mitigation methods, such as self-reflection and decoding strategies, correct generated content but often introduce additional steps, increasing both cost and latency (Anonymous, 2024c; Arteaga et al., 2024). In contrast, data- and training-level methods refine the model using specialized datasets and alignment algorithms like RLHF and DPO. However, they share similar challenges, requiring substantial data and computational resources (Xing et al., 2024; Rawte et al., 2023). Therefore, future work should not only focus on hallucination mitigation but also explore strategies for improving efficiency.

Explainability. While research has focused on detecting hallucinations by comparing outputs to factual data, the underlying mechanisms driving these hallucinations remain underexplored. Studies examining model confidence through logits provide some insights (Hou et al., 2024; Valentin et al., 2024), but a deeper understanding of how specific attention heads or neuron activations contribute to hallucination mitigation is still needed. Some work has explored decoding strategies (Chen et al., 2024a), and model editing techniques (Xu et al., 2024a; Mishra et al., 2024) may emerge as promising interpretability-oriented solutions for combating medical hallucination in future research.

Multilinguality. Current research has primarily focused on English. Some studies have begun to explore hallucination in the general domain within LLMs (Chataigner et al., 2024) and LVLMs (Qu et al., 2024). Future research should expand to address medical hallucination in non-English texts, leveraging multilingual LLMs (Jiang et al., 2023; Ming et al., 2024) and LVLMs (Li et al., 2023d; Chen et al., 2024d) to account for language-specific characteristics. This will enhance medical accuracy and accessibility in low-resource language regions.

8 Discussion on Quantitative Analysis

While quantitative analysis of medical hallucinations in LLMs and LVLMs, including the impact and comparative performance of mitigation strategies, would be valuable, such an endeavor faces significant challenges. The complexity of conducting new, large-scale quantitative experiments is considerable, particularly given the disparate datasets used in existing behavioral analyses and the variability in evaluation baselines (e.g., different model versions), as detailed in Tables 2, 6 for LLMs and Tables 3, 7 for LVLMs. Consequently, establishing and maintaining unified, comprehensive evaluation benchmarks and metrics for fair comparison remains a crucial direction for future research.

Nevertheless, this survey synthesizes existing quantitative results on mitigation strategies by compiling findings from various studies, presented in Table 4 for LLMs and Table 5 for LVLMs.

9 Conclusion

In this paper, we conduct a systematic survey of medical hallucination in both LLMs and LVLMs. Concretely, we meticulously categorize medical hallucination into a unified perspective and trace recent studies from the aspects of causes, detection, evaluation, and mitigation. Moreover, we delineate the challenges and delve into future frontiers. We hope that this survey will facilitate further research in medical hallucination.

Limitations

We have made our best effort, but there may still be some limitations. On one hand, due to page limitations, we can only provide a brief summary of each method without exhaustive technical details. On the other hand, we primarily collect studies from *ACL, NeurIPS, ICLR, ICML and arXiv, etc. As such, there is a chance that we may have missed some important work published in other venues. We will stay abreast of discussions within the research community, updating opinions and supplementing overlooked work in the future.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12.

Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 25–29. IOS Press.

Angus Addlesee. 2024. Grounding llms to in-prompt instructions: Reducing hallucinations caused by static pre-training knowledge. In *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI@ LREC-COLING 2024*, pages 1–7.

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. Medhalu: Hallucinations in responses to healthcare queries by large language models. *arXiv preprint arXiv:2409.19492*.

Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. 2023. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*.

Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A Mesallam, Mohamed Farahat, Khalid H Malki, et al. 2024. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Medical Informatics*, 12(1):e54345.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Sumera Anjum, Hanzhi Zhang, Wenjun Zhou, Eun Jin Paek, Xiaopeng Zhao, and Yunhe Feng. 2024. Halo: Hallucination analysis and learning optimization to empower llms with retrieval-augmented context for guided clinical decision making. *arXiv preprint arXiv:2409.10011*.

Anonymous. 2024a. Hademif: Hallucination detection and mitigation in large language models. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Anonymous. 2024b. Hallucination mitigating for medical report generation. In *Submitted to ACL Rolling Review - June 2024*.

Anonymous. 2024c. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Gabriel Y Arteaga, Thomas B Schön, and Nicolas Pielawski. 2024. Hallucination detection in llms: Fast and memory-efficient finetuned models. *arXiv preprint arXiv:2409.02976*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. 2023c. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6859–6865. IEEE.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Oishi Banerjee, Hong-Yu Zhou, Subathra Adithan, Stephen Kwak, Kay Wu, and Pranav Rajpurkar. 2024. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. *arXiv preprint arXiv:2406.06496*.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021.

Gauri Bhardwaj, Yuvaraj Govindarajulu, Sundaraparipurnan Narayanan, Pavan Kulkarni, and Manojkumar Parmar. 2023. On the notion of hallucinations from the lens of bias and validity in synthetic cxr images. *arXiv preprint arXiv:2312.06979*.

Junyi Bian, Jiaxuan Zheng, Yuyi Zhang, and Shanfeng Zhu. 2023. Inspire the large language model by external knowledge on biomedical named entity recognition. *arXiv preprint arXiv:2309.12278*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark A Burgess, Brendan Hosking, Roc Reguant, Anubhav Kaphle, Mitchell J O'Brien, Letitia MF Sng, Yatish Jain, and Denis C Bauer. 2024. Privacy-hardened and hallucination-resistant synthetic data generation with logic-solvers. *arXiv preprint arXiv:2410.16705*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2024. A survey on mixture of experts. *Authorea Preprints*.

Shuyang Cao and Lu Wang. 2021. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cléa Chataigner, Afaf Taïk, and Golnoosh Farnadi. 2024. Multilingual hallucination gaps in large language models. *arXiv preprint arXiv:2410.18270*.

Beitao Chen, Xinyu Lyu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. 2024a. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. 2023a. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*.

Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024b. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023b. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. 2024c. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023c. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. 2024d. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14432–14444.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. 2024e. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023d. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024f. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. 2024g. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024h. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Tugba Akinci D'Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. 2024. Totalsegmentator mri: Sequence-independent segmentation of 59 anatomical structures in mr images. *arXiv preprint arXiv:2405.19492*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Chengfeng Dou, Ying Zhang, Yanyuan Chen, Zhi Jin, Wenpin Jiao, Haiyan Zhao, and Yu Huang. 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)*, pages 4784–4794.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

Nima Ebadi, Kellen Morgan, Adrian Tan, Billy Linares, Sheri Osborn, Emma Majors, Jeremy Davis, and Anthony Rios. 2024. Extracting biomedical entities from noisy audio transcripts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7023–7034, Torino, Italia. ELRA and ICCL.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*.

Zishan Gu, Changchang Yin, Fenglin Liu, and Ping Zhang. 2024. Medvh: Towards systematic evaluation of hallucination for large vision language models in the medical context. *CoRR*.

Danfeng Guo and Demetri Terzopoulos. 2024. Prompting medical large vision-language models to diagnose pathologies by visual question answering. *arXiv preprint arXiv:2407.21368*.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023a. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Xinzhe Han, Shuhui Wang, Chi Su, Qingming Huang, and Qi Tian. 2023b. General greedy de-bias learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9789–9805.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020a. Pathological visual question answering. *arXiv preprint arXiv:2010.12435*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020b. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Stefan Hegselmann, Shannon Zejiang Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. 2024. A data-centric approach to generate faithful and high quality patient summaries with large language models. *arXiv preprint arXiv:2402.15422*.

Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. 2024. A probabilistic framework for llm hallucination detection via belief tree propagation. *arXiv preprint arXiv:2406.06950*.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. 2023. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4156–4165.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023b. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Minbyul Jeong, Hyeon Hwang, Chanwoong Yoon, Taewhoo Lee, and Jaewoo Kang. 2024a. Olaph: Improving factuality in biomedical long-form question answering. *arXiv preprint arXiv:2405.12701*.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024b. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2024. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. *Preprint*, arXiv:2406.11451.

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

Liqiang Jing and Xinya Du. 2024. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019a. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019b. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Praveen K Kanithi, Clément Christophe, Marco AF Pimentel, Tathagata Raha, Nada Saadi, Hamza Javed, Svetlana Maslenkova, Nasir Hayat, Ronnie Rajan, and Shadab Khan. 2024. Medic: Towards a comprehensive framework for evaluating llms in clinical applications. *arXiv preprint arXiv:2409.07314*.

Yuhe Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Yilin Ning, Irene Li, Hairil Rizal Abdullah, Daniel Shu Wei Ting, and Nan Liu. 2024. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: Simulation study. *Journal of Medical Internet Research*, 26:e59439.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. 2024. Generating synthetic data for medical imaging. *Radiology*, 312(3):e232471.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. 2023. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. *arXiv preprint arXiv:2305.11490*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. 2024b. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Songda Li, Yunqi Zhang, Chunyuan Deng, Yake Niu, and Hui Zhao. 2024c. Better late than never: Model-agnostic hallucination post-processing framework towards clinical text summarization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 995–1011.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuanjing Huang, and Zhongyu Wei. 2023d. Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5939–5958, Toronto, Canada. Association for Computational Linguistics.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Zhou, and Zuozhu Liu. 2024d. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024e. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993*.

Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. 2024a. Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint arXiv:2401.08358*.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. 2024b. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301.

Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13151–13160.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-QA: A real-world medical Q&A benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 277–294, Bangkok, Thailand. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, et al. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. *arXiv preprint arXiv:2412.04003*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *arXiv preprint arXiv:2401.06855*.

Prakamya Mishra, Zonghai Yao, Shuwei Chen, Beining Wang, Rohan Mittal, and Hong Yu. 2023. Synthetic imitation edit feedback for factual alignment in clinical summarization. *arXiv preprint arXiv:2310.20033*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *arXiv preprint arXiv:2305.15852*.

Yang Nan, Huichi Zhou, Xiaodan Xing, and Guang Yang. 2024. Beyond the hype: A dispassionate look at vision-language models in medical scenario. *arXiv preprint arXiv:2408.08704*.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What lms know about unseen entities. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 693–702.

OpenAI. 2022. ChatGPT blog. https://openai.com/blog/chatgpt.

OpenAI et al. 2023. GPT-4 technical report. *arXiv Preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Duy Khoa Pham and Bao Quoc Vo. 2024. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. *arXiv preprint arXiv:2408.13808*.

Lang Qin, Yao Zhang, Hongru Liang, Adam Jatowt, and Zhenglu Yang. 2024. Listen to the patient: Enhancing medical dialogue generation with patient hallucination detection and mitigation. *arXiv preprint arXiv:2410.06094*.

Xiaoye Qu, Mingyang Song, Wei Wei, Jianfeng Dong, and Yu Cheng. 2024. Mitigating multilingual hallucination in large vision-language models. *arXiv preprint arXiv:2408.00550*.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Sudha Rao. 2017. Are you asking the right questions? teaching machines to ask clarification questions. In *Proceedings of ACL 2017, Student Research Workshop*, pages 30–35.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Corentin Royer, bjoern menze, and Anjany Sekuboyina. 2024. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. In *Medical Imaging with Deep Learning*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724.

Sraavya Sambara, Serena Zhang, Oishi Banerjee, Julian Acosta, John Fahrner, and Pranav Rajpurkar. 2024. Radflag: A black-box hallucination detection method for medical vision language models. *arXiv preprint arXiv:2411.00299*.

Gordon D Schiff, Omar Hasan, Seijeoung Kim, Richard Abrams, Karen Cosby, Bruce L Lambert, Arthur S Elstein, Scott Hasler, Martin L Kabongo, Nela Krosnjar, et al. 2009. Diagnostic error in medicine: analysis of 583 physician-reported errors. *Archives of internal medicine*, 169(20):1881–1887.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A Smith, Luke Zettlemoyer, et al. 2023. In-context pretraining: Language modeling beyond document boundaries. *arXiv preprint arXiv:2310.10638*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8.

Eduardo Soares, Plamen Angelov, Sarah Biaso, Michele Higa Froes, and Daniel Kanda Abe. 2020. Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. *MedRxiv*, pages 2020–04.

Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. 2020. Medicat: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2112–2120.

Li Sun, Liuan Wang, Jun Sun, and Takayuki Okatani. 2025. Temporal insight enhancement: Mitigating temporal hallucination in video understanding by multimodal large language models. In *International Conference on Pattern Recognition*, pages 455–473. Springer.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13088–13110.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. MedAgents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621, Bangkok, Thailand. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023a. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

MosaicML NLP Team et al. 2023b. Introducing mpt-30b: Raising the bar for open-source foundation models.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Simon Valentin, Jinmiao Fu, Gianluca Detommaso, Shaoyuan Xu, Giovanni Zappella, and Bryan Wang. 2024. Cost-effective hallucination detection for llms. *arXiv preprint arXiv:2407.21424*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

David Vilares and Carlos Gómez-Rodríguez. 2019. Head-qa: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966.

Prathiksha Rumale Vishwanath, Simran Tiwari, Tejas Ganesh Naik, Sahil Gupta, Dung Ngoc Thai, Wenlong Zhao, SUNJAE KWON, Victor Ardulov, Karim Tarabishy, Andrew McCallum, et al. 2024. Faithfulness hallucination detection in healthcare ai. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, et al. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*.

Huimin Wang, Yutian Zhao, Xian Wu, and Yefeng Zheng. 2024a. imapscore: Medical fact evaluation made easy. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10242–10257.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024b. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024c. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. 2023. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Jinge Wu, Yunsoo Kim, and Honghan Wu. 2024a. Hallucination benchmark in medical visual question answering. In *The Second Tiny Papers Track at ICLR 2024*.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024b. Evaluating and analyzing relationship hallucinations in large vision-language models. *arXiv preprint arXiv:2406.16449*.

Xian Wu, Yutian Zhao, Yunyan Zhang, Jiageng Wu, Zhihong Zhu, Yingying Zhang, Yi Ouyang, Ziheng Zhang, Huimin WANG, Zhenxi Lin, Jie Yang, Shuang Zhao, and Yefeng Zheng. 2024c. Medjourney: Benchmark and evaluation of large language models over patient clinical journey. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024d. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*.

Peng Xia, Ze Chen, Juanxi Tian, Gong Yangrui, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, Wenhao Zheng, Zhaoyang Wang, Xiao Wang, Xuchao Zhang, Chetan Bansal, Marc Niethammer, Junzhou Huang, Hongtu Zhu, Yun Li, Jimeng Sun, Zongyuan Ge, Gang Li, James Zou, and Huaxiu Yao. 2024a. CARES: A comprehensive benchmark of trustworthiness in medical vision language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b. Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093.

Tong Xiang, Liangzhi Li, Wangyue Li, Mingbai Bai, Lu Wei, Bowen Wang, and Noa Garcia. 2023. Caremi: Chinese benchmark for misinformation evaluation in maternity and infant care. In *Advances in Neural Information Processing Systems*, volume 36, pages 42358–42381. Curran Associates, Inc.

Yong Xie, Karan Aggarwal, Aitzaz Ahmad, and Stephen Lau. 2024. Controlled automatic task-specific synthetic data generation for hallucination detection. *arXiv preprint arXiv:2410.12278*.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2024. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2402.09801*.

Xiaodan Xing, Giorgos Papanastasiou, Simon Walsh, and Guang Yang. 2023. Less is more: unsupervised mask-guided annotated ct image synthesis with minimum manual segmentations. *IEEE Transactions on Medical Imaging*, 42(9):2566–2576.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Wanyu Wang, Yuyang Ye, Xiangyu Zhao, et al. 2024a. Editing factual knowledge and explanatory ability of medical large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2660–2670.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024b. Mitigating hallucinations of large language models in medical information extraction via contrastive decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7744–7757.

Qianqi Yan, Xuehai He, and Xin Eric Wang. 2024a. Med-hvl: Automatic medical domain hallucination evaluation for large vision-language models. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2024b. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *arXiv preprint arXiv:2405.20421*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*.

Juan M Zambrano Chaves, Andrew L Wentland, Arjun D Desai, Imon Banerjee, Gurkiran Kaur, Ramon Correa, Robert D Boutin, David J Maron, Fatima Rodriguez, Alexander T Sandhu, et al. 2023. Opportunistic assessment of ischemic heart disease risk using abdominopelvic computed tomography and medical record data: a multimodal explainable artificial intelligence approach. *Scientific reports*, 13(1):21034.

Dongxu Zhang, Varun Gangal, Barrett Martin Lattimer, and Yi Yang. 2024a. Enhancing hallucination detection through perturbation-based synthetic data generation in system responses. *arXiv preprint arXiv:2407.05474*.

Fan Zhang, Hao Chen, Zhihong Zhu, Ziheng Zhang, Zhenxi Lin, Ziyue Qiao, Yefeng Zheng, and Xian Wu. 2025a. A survey on foundation language models for single-cell biology. In *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Fan Zhang, Tianyu Liu, Zhihong Zhu, Hao Wu, Haixin Wang, Donghao Zhou, Yefeng Zheng, Kun Wang, Xian Wu, and Pheng-Ann Heng. 2025b. Cellverse: Do large language models really understand cell biology? *arXiv preprint arXiv:2505.07865*.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. 2024b. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141.

Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang. 2023a. Famesumm: Investigating and improving faithfulness of medical summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10915–10931.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023c. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Jiawei Zhu, Yishu Liu, Huanjia Zhu, Hui Lin, Yuncheng Jiang, Zheng Zhang, and Bingzhi Chen. 2024. Combating visual question answering hallucinations via robust multi-space co-debias learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 955–964.

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K Reddy. 2019. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, pages 2472–2482.

Reference	Model	Dataset	Application	Quantitative Results
FaMeSumm (Zhang et al., 2023a)	PEGASUS, BART, BioBART, T5, mT5	HQS, RRS, MDS	Health Question Summarization, Radiology Report Summarization, Medical Dialogue Summarization	FAMESUMM generates 16% more faithful summaries than GPT-3 based on doctors' evaluation, and provides consistent score improvements over baselines according to automatic metrics.
Self-Reflection (Ji et al., 2023b)	Vicuna, Alpaca-LoRA, ChatGPT, MedAlpaca, Robin-medical	PubMedQA, MedQuAD, MEDIQA2019, LiveMedQA2017, MASH-QA	Medical Natural Language Inference	Alpaca-LoRA-7B with the proposed method gains around 3× larger improvement than baseline for Sample- and Sentence-level MedNLI on PubMedQA.
MedPH (Qin et al., 2024)	DISC-MedLLM, HuatuoGPT2, GPT-3.5, GPT-4, GPT-2, VIBBot, DFMED, EMULATION	MedDG, KaMed	Medical Dialogue Systems	The proposed method improves performance by up to 10% on entity prediction and response generation tasks.
OLAPH (Jeong et al., 2024a)	Llama-2, Mistral, Meditron, BioMistral, Self-BioRAG, GPT-3.5, Claude 3 Sonnet, GPT-4o	LiveQA, MedicationQA, HealthSearchQA, K-QA	Biomedical Long-form Question Answering	After three iterations of DPO training, the responses from BioMistral 7B approach GPT-4-level performance.
MEDAL (Li et al., 2024c)	PEGASUS, BioBART, Llama-2, Med-Alpaca	HQS, RRS, ACI-BENCH	Health Question Summarization, Radiology Report Summarization, Doctor-Patient Dialogue Summarization	The method improves Llama-2's performance by 6% on health question summarization and enhances Med-Alpaca's performance by 9% on radiology report summarization.
ALCD (Xu et al., 2024b)	ChatGLM-6B, Qwen-7B-Chat	CMEE-V2, CMEIE-V2, IMCS-V2-NER, CMedCausal, IMCS-V2-SR, CHIP-MDCFNPC	Medical Information Extraction	ALCD outperforms previous decoding methods, with the largest performance gap reaching 4.50% for Qwen-7B-chat on the CMedCausal dataset.
HALO (Anjum et al., 2024)	ChatGPT 3.5, Llama-3.1, Mistral	MedMCQA	Medical Question Answering	HALO improves ChatGPT's accuracy from 44% to 65% on the TEST subset.

Table 4: Comparison of quantitative results on the effectiveness of various mitigation strategies in LLMs.

Reference	Model	Dataset	Application	Quantitative Results
CoMT (Jiang et al., 2024)	LLaVA-Med, MiniGPT4, XrayGPT, mPLUG-Owl2, R2Gen	OpenI, MIMIC-CXR	Medical Report Generation	Models trained using CoMT data show improvements in NLG metrics and surpass models trained with original MRG data by 2%–5% in hallucination metrics.
RULE (Xia et al., 2024b)	LLaVA-Med-1.5	IU-Xray, Harvard-FairVLMed, MIMIC-CXR	Medical Visual Question Answering, Medical Report Generation	RULE achieves 47.4% average accuracy improvement on two tasks across all datasets.
MedCoT (Liu et al., 2024d)	MEVF, MMBERT, PubMedCLIP, VQA-Adapter, MedThink, LLaVA-Med	VQA-RAD, SLAKE-EN, Med-VQA-2019, PathVQA	Medical Visual Question Answering	MedCoT outperforms Gemini by 27.21% on VQA-RAD and 14.66% on SLAKE-EN.
KERM (Anonymous, 2024b)	R2Gen, HRGR, CoAtt, PKERRG, CMAS-RL, CMM, CCR, PPKED, KM, Multicriteria	IU-Xray, MIMIC-CXR	Medical Report Generation	KERM achieves BLEU-4: 0.182, METEOR: 0.197, ROUGE-L: 0.388 on IU-Xray dataset.
DPO-RRG (Banerjee et al., 2024)	Swin Transformer + Llama2-Chat-7b	MIMIC-CXR	Medical Report Generation	DPO reduces reports mentioning prior exams to 20%–25%, halving the original proportion.

Table 5: Comparison of quantitative results on the effectiveness of various mitigation strategies in LVLMs.

Benchmark	Evaluation Baseline	Underlying Data Source
Med-HALT (Pal et al., 2023)	Text Davinci (Brown et al., 2020) GPT-3.5 Llama-2 (Touvron et al., 2023b) MPT (Team et al., 2023b) Falcon (Penedo et al., 2023)	MedMCQA (Pal et al., 2022) HEAD-QA (Vilares and Gómez-Rodríguez, 2019) MedQA (Jin et al., 2021) PubMed
MedHalu (Agarwal et al., 2024)	Llama-2 (Touvron et al., 2023b) GPT-3.5 GPT-4 (OpenAI et al., 2023) Human	HealthQA (Zhu et al., 2019) LiveQA (Abacha et al., 2017) MedicationQA (Abacha et al., 2019)
CMHE-HD (Dou et al., 2024)	ChatGPT Baichuan (Yang et al., 2023) Qwen (Bai et al., 2023a)	CMD[4] cMedQA2 (Zhang et al., 2018)
K-QA (Manes et al., 2024)	Mistral (Jiang et al., 2023) MedAlpaca (Han et al., 2023a) LLama (Touvron et al., 2023a) GPT-3.5 GPT-4 (OpenAI et al., 2023) PALM-2 (Anil et al., 2023) BARD[6] Bing Chat[7]	K Health[5]
MedLFQA (Jeong et al., 2024a)	LLama-2 (Touvron et al., 2023b) Mistral (Jiang et al., 2023) Meditron (Chen et al., 2023d) BioMistral (Labrak et al., 2024) Self-BioRAG (Jeong et al., 2024b) GPT-3.5 Claude 3 Sonnet[8] GPT-4o (Hurst et al., 2024)	LiveQA (Abacha et al., 2017) MedicationQA (Abacha et al., 2019) HealthSearchQA (Singhal et al., 2023) K-QA (Manes et al., 2024)
*imap*Bench (Wang et al., 2024a)	GPT-4 (OpenAI et al., 2023) ChatGPT PALM-2 (Anil et al., 2023)	HMedQA[9] iCliniq[10]

[4] https://github.com/Toyhom/Chinese-medical-dialogue-data [5] https://khealth.com/
[6] https://bard.google.com/ [7] https://www.bing.com/ [8] https://claude.ai/
[9] https://github.com/SCIR-HI/Huatuo-Llama-Med-Chinese/blob/main/data/llama_data.json
[10] https://github.com/Kent0n-Li/ChatDoctor?tab=readme-ov-file#resources-list

Table 6: Details of medical hallucination benchmark for LLMs.

Benchmark	Evaluation Baseline	Underlying Data Source
MedVH (Gu et al., 2024)	GPT-4 (OpenAI et al., 2023) MiniGPT (Chen et al., 2023b) LLaVA (Liu et al., 2024c) LLaVA-Med (Li et al., 2024a) Med-Flamingo (Moor et al., 2023) CheXAgent (Chen et al., 2024g) LLM-CXR (Lee et al., 2023)	VQA-RAD (Lau et al., 2018) SLAKE (Liu et al., 2021) PMC-VQA (Zhang et al., 2023b) Path-VQA (He et al., 2020b) VQA-Med-2021 (Ben Abacha et al., 2021) MIMIC-Diff-VQA (Hu et al., 2023)
Halt-MedVQA (Wu et al., 2024a)	LLaVA (Liu et al., 2024c) LLaVA-Med (Li et al., 2024a) GPT-4 (OpenAI et al., 2023)	PMC-VQA (Zhang et al., 2023b) PathVQA (He et al., 2020b) VQA-RAD (Lau et al., 2018)
Med-HallMark (Chen et al., 2024b)	BLIP2 (Li et al., 2023a) InstructBLIP (Dai et al., 2023) LLaVA1.5 (Liu et al., 2024c) mPLUG-Owl2 (Ye et al., 2024) XrayGPT (Thawkar et al., 2023) MiniGPT4 (Zhu et al., 2023) RadFM (Wu et al., 2023) LLaVA-Med (Li et al., 2024a)	SLAKE (Liu et al., 2021) VQA-RAD (Lau et al., 2018) MIMIC (Johnson et al., 2019a) OpenI (Wang et al., 2017)
CARES-Fact (Xia et al., 2024a)	Qwen-VL-Chat (Bai et al., 2023b) LLaVA-1.6 (Liu et al., 2024b) LLaVA-Med (Li et al., 2024a) Med-Flamingo (Moor et al., 2023) RadFM (Wu et al., 2023) MedVInT (Zhang et al., 2023b)	MIMIC-CXR (Johnson et al., 2019b) IU-Xray (Demner-Fushman et al., 2016) Harvard-FairVLMed (Luo et al., 2024b) PMC-OA (Lin et al., 2023) HAM10000 (Tschandl et al., 2018) OL3I (Zambrano Chaves et al., 2023) OmniMedVQA (Hu et al., 2024)
RadVUQA (Nan et al., 2024)	LLaVA (Liu et al., 2024c) InternVL (Chen et al., 2024f) MiniCPM (Yao et al., 2024) BLIP2 (Li et al., 2023a) LLaVA-Med (Li et al., 2024a) HuatuoGPT-Vision (Chen et al., 2024c) GPT-4o (Hurst et al., 2024)	RadVUQA-CT (Wasserthal et al., 2023) RadVUQA-MRI (D'Antonoli et al., 2024) RadVUQA-OOD-1 (Xing et al., 2023) RadVUQA-OOD-2 (Soares et al., 2020) RadVUQA-OOD-3[11]
ProbMed (Yan et al., 2024b)	GPT-4o (Hurst et al., 2024) Gemini Pro (Team et al., 2023a) LLaVA (Liu et al., 2024b) MiniGPT-v2 (Chen et al., 2023b) LLaVA-Med (Li et al., 2024a) Med-Flamingo (Moor et al., 2023) BiomedGPT (Zhang et al., 2024b) RadFM (Wu et al., 2023) CheXAgent (Chen et al., 2024g) GPT-4v[12]	MedICaT (Subramanian et al., 2020) ChestX-ray14 (Wang et al., 2017)

[11] https://www.embodi3d.com/ [12] https://openai.com/index/gpt-4v-system-card/

Table 7: Details of medical hallucination benchmark for LVLMs.