A Comprehensive Survey on the Trustworthiness of Large Language Models in Healthcare

Anonymous ACL submission

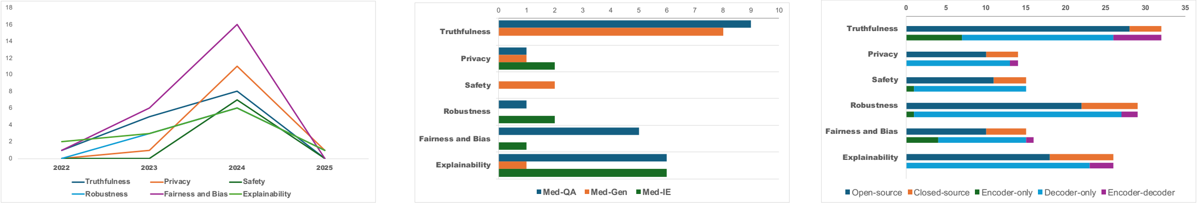
Abstract

The application of large language models (LLMs) in healthcare holds significant promise for enhancing clinical decision-making, medical research, and patient care. However, their integration into real-world clinical settings raises critical concerns around trustworthiness, particularly around dimensions of truthfulness, privacy, safety, robustness, fairness, and explainability. These dimensions are essential for ensuring that LLMs generate reliable, unbiased, and ethically sound outputs. While researchers have recently begun developing benchmarks and evaluation frameworks to assess LLM trustworthiness, the **trustworthiness of LLMs in healthcare** remains underexplored, lacking a systematic review that provides a comprehensive understanding and future insights. This **survey** addresses that gap by providing a comprehensive review of current methodologies and solutions aimed at mitigating risks across key trust dimensions. We analyze how each dimension affects the reliability and ethical deployment of healthcare LLMs, synthesize ongoing research efforts and identify critical gaps in existing approaches. We also identify emerging challenges posed by evolving paradigms, such as multi-agent collaboration, multi-modal reasoning, and the development of small open-source medical models. Our goal is to guide future research toward more trustworthy, transparent, and clinically viable LLMs.

1 Introduction

The application of LLMs in healthcare is advancing rapidly, with the potential to transform clinical decision-making, medical research, and patient care. However, incorporating them into healthcare systems poses several key challenges that need to be addressed to ensure their reliable and ethical use. As highlighted in Bi et al. (2024), a major concern is the trustworthiness of AI-enhanced biomedical insights. This encompasses improving model explainability and interpretability, enhancing robustness against adversarial attacks, mitigating biases across diverse populations, and ensuring strong data privacy protections. Key concerns include truthfulness, privacy, safety, robustness, fairness, and explainability, each of which plays a vital role in the reliability and trustworthiness of AI-driven healthcare solutions.

Truthfulness, defined as "the accurate representation of information, facts, and results by an AI system" (Huang et al., 2024), is critical in healthcare, as inaccuracies can lead to misdiagnoses or inappropriate treatment recommendations. Ensuring that generated information is both accurate and aligned with verified medical knowledge is essential. Additionally, *privacy* concerns arise from the risk of exposing sensitive patient data during model training and usage, potentially leading to breaches or violations of regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation). Ensuring patient confidentiality while leveraging LLMs for diagnostics and treatment recommendations is a critical challenge. *Safety*, defined as “ensuring that LLMs do not answer questions that can harm patients or healthcare providers in healthcare settings” (Han et al., 2024b), further underscores the necessity of implementing stringent safeguards to mitigate harm. *Robustness* refers to an LLM’s ability to consistently generate accurate, reliable, and unbiased outputs across diverse clinical scenarios while minimizing errors, hallucinations, and biases. It also encompasses the model’s resilience against adversarial attacks, ensuring that external manipulations do not compromise its integrity. A truly robust LLM in healthcare must demonstrate stability, reliability, and fairness, even when faced with noisy, ambiguous, or adversarial inputs. Similarly, *fairness and bias* must be addressed to prevent discriminatory



(a) Temporal Trends

(b) Distribution of Datasets

(c) Distribution of Models

Figure 1: Overview of research trends, dataset usage, and model types across key trustworthiness dimensions in healthcare LLM studies: (a) Temporal Trends in Trustworthiness Dimensions Addressed in Medical LLM Studies (2022–2025); (b) Distribution of Dataset Types Across Trustworthiness Dimensions in Healthcare LLM Studies; (c) Distribution of Model Types Across Trustworthiness Dimensions in Healthcare LLM Studies.

patterns in model outputs, which could lead to unequal treatment recommendations and exacerbate healthcare disparities. Furthermore, the *explainability* of LLMs, which ensures that model outputs are interpretable and transparent, plays a vital role in fostering trust and allowing informed decision-making by healthcare professionals. Lack of transparency in model reasoning complicates clinical adoption and raises accountability concerns.

Tackling these challenges is essential for the trustworthy and ethical implementation of LLMs in healthcare. Recently, researchers have begun developing benchmarks and evaluation frameworks to systematically assess the trustworthiness of LLMs (Huang et al., 2024). The **trustworthiness of LLMs in healthcare** is gaining increasing attention due to its significant social impact. However, there is currently no systematic review that provides a comprehensive understanding and future insights into this area. To bridge this gap, we present a comprehensive **survey** that explores these trust-related dimensions in detail, reviewing existing benchmarks, and methodologies aimed at improving the trustworthiness of LLMs in healthcare.

2 Datasets, Models, and Tasks

2.1 Inclusion & Exclusion Criteria

We initiated our survey with a comprehensive literature search targeting studies on the trustworthiness of LLMs in healthcare. Our search strategy employed diverse keyword combinations and was directed toward top-tier conferences and journals, prioritizing publications from 2022 onward. Detailed inclusion and exclusion criteria are provided in Appendix A. Fig 1(a) illustrates how the number of papers addressing each key trustworthiness dimension in healthcare LLMs has changed over time from 2022 to 2025. From Figure 1(a), interest in trustworthiness dimensions peaked in 2024, particularly for Fairness and Bias (16 papers) and Privacy (11 papers), reflecting a strong recent push toward ethical and secure AI in healthcare. Truthfulness and Explainability maintained steady growth through 2023 and 2024. These trends suggest a rising concern with fairness and privacy in recent years, possibly driven by real-world deployment risks and regulatory pressure.

2.2 Datasets

The datasets used in studies of trust in LLMs for healthcare are categorized by the dimensions of trustworthiness they address in Appendix B, where we highlight key details such as data type, content, task, and dimensions of trustworthiness. The content of each dataset specifies its composition, while the task refers to the main problem to be solved for which the dataset is utilized. The data type varies across studies and includes web-scraped data, curated domain-specific datasets, public text corpora, synthetic data, real-world data, and private datasets, providing a comprehensive overview of their relevance to healthcare applications.

Figure 1(b) shows the number of studies using three major dataset types—Med-QA (blue), Med-Gen (orange), and Med-IE (green)—in relation to six trustworthiness dimensions: Truthfulness, Privacy, Safety, Robustness, Fairness and Bias, and Explainability. Figure 1(b) shows how three major dataset types—Med-QA, Med-Gen, and Med-IE—are used across six trust dimensions. Truthfulness is most studied with both Med-QA and Med-Gen. Med-QA is also common in fairness and explainability, while Med-Gen contributes to safety and privacy. Med-IE, though less used overall, is more prominent in robustness and explainability. This highlights the dominance of Med-QA

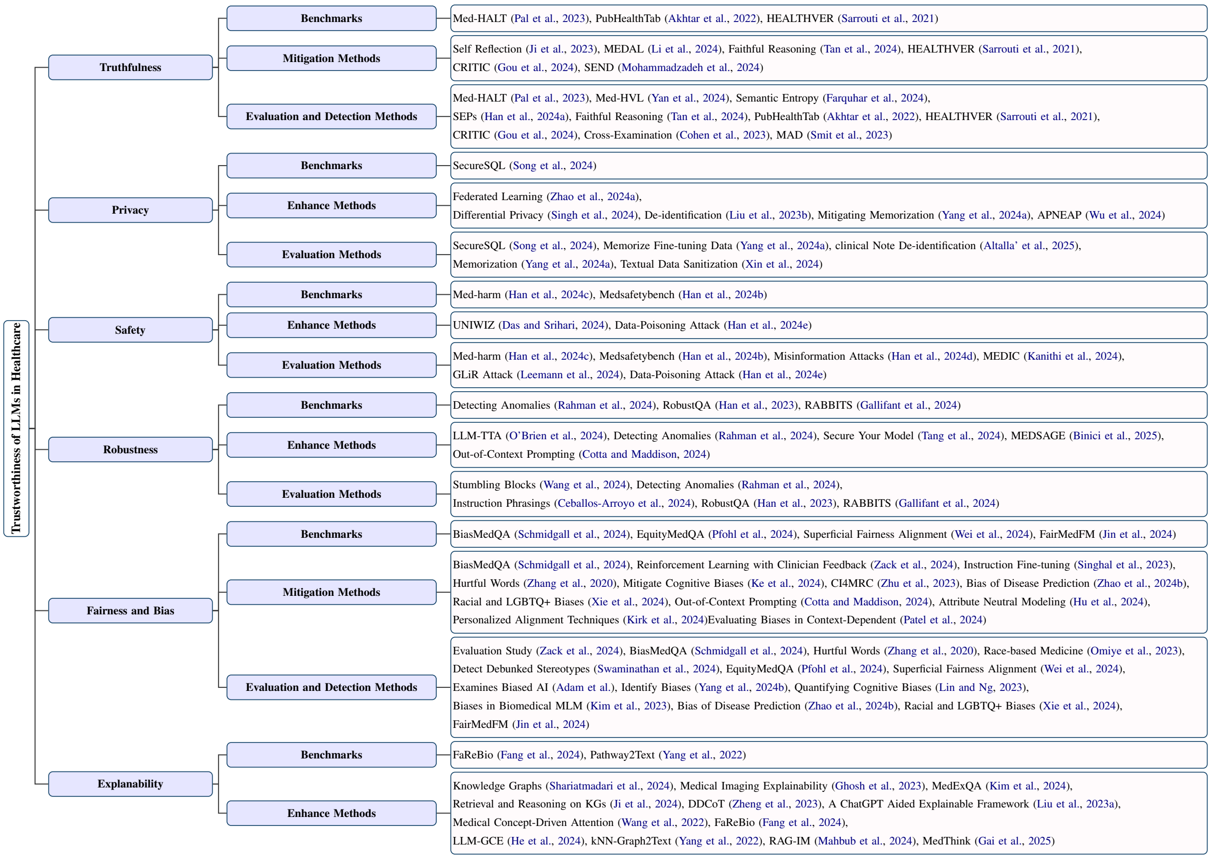

Trustworthiness of LLMs in Healthcare
Truthfulness
Benchmarks
Med-HALT (Pal et al., 2023), PubHealthTab (Akhtar et al., 2022), HEALTHVER (Sarrouti et al., 2021)
Mitigation Methods
Self Reflection (Ji et al., 2023), MEDAL (Li et al., 2024), Faithful Reasoning (Tan et al., 2024), HEALTHVER (Sarrouti et al., 2021), CRITIC (Gou et al., 2024), SEND (Mohammadzadeh et al., 2024)
Evaluation and Detection Methods
Med-HALT (Pal et al., 2023), Med-HVL (Yan et al., 2024), Semantic Entropy (Farquhar et al., 2024), SEPs (Han et al., 2024a), Faithful Reasoning (Tan et al., 2024), PubHealthTab (Akhtar et al., 2022), HEALTHVER (Sarrouti et al., 2021), CRITIC (Gou et al., 2024), Cross-Examination (Cohen et al., 2023), MAD (Smit et al., 2023)
Privacy
Benchmarks
SecureSQL (Song et al., 2024)
Enhance Methods
Federated Learning (Zhao et al., 2024a), Differential Privacy (Singh et al., 2024), De-identification (Liu et al., 2023b), Mitigating Memorization (Yang et al., 2024a), APNEAP (Wu et al., 2024)
Evaluation Methods
SecureSQL (Song et al., 2024), Memorize Fine-tuning Data (Yang et al., 2024a), clinical Note De-identification (Altalla' et al., 2025), Memorization (Yang et al., 2024a), Textual Data Sanitization (Xin et al., 2024)
Safety
Benchmarks
Med-harm (Han et al., 2024c), Medsafetybench (Han et al., 2024b)
Enhance Methods
UNIWIZ (Das and Srihari, 2024), Data-Poisoning Attack (Han et al., 2024e)
Evaluation Methods
Med-harm (Han et al., 2024c), Medsafetybench (Han et al., 2024b), Misinformation Attacks (Han et al., 2024d), MEDIC (Kanithi et al., 2024), GLiR Attack (Leemann et al., 2024), Data-Poisoning Attack (Han et al., 2024e)
Robustness
Benchmarks
Detecting Anomalies (Rahman et al., 2024), RobustQA (Han et al., 2023), RABBITS (Gallifant et al., 2024)
Enhance Methods
LLM-TTA (O'Brien et al., 2024), Detecting Anomalies (Rahman et al., 2024), Secure Your Model (Tang et al., 2024), MEDSAGE (Binici et al., 2025), Out-of-Context Prompting (Cotta and Maddison, 2024)
Evaluation Methods
Stumbling Blocks (Wang et al., 2024), Detecting Anomalies (Rahman et al., 2024), Instruction Phrasings (Ceballos-Arroyo et al., 2024), RobustQA (Han et al., 2023), RABBITS (Gallifant et al., 2024)
Fairness and Bias
Benchmarks
BiasMedQA (Schmidgall et al., 2024), EquityMedQA (Pfohl et al., 2024), Superficial Fairness Alignment (Wei et al., 2024), FairMedFM (Jin et al., 2024)
Mitigation Methods
BiasMedQA (Schmidgall et al., 2024), Reinforcement Learning with Clinician Feedback (Zack et al., 2024), Instruction Fine-tuning (Singhal et al., 2023), Hurtful Words (Zhang et al., 2020), Mitigate Cognitive Biases (Ke et al., 2024), CI4MRC (Zhu et al., 2023), Bias of Disease Prediction (Zhao et al., 2024b), Racial and LGBTQ+ Biases (Xie et al., 2024), Out-of-Context Prompting (Cotta and Maddison, 2024), Attribute Neutral Modeling (Hu et al., 2024), Personalized Alignment Techniques (Kirk et al., 2024)Evaluating Biases in Context-Dependent (Patel et al., 2024)
Evaluation and Detection Methods
Evaluation Study (Zack et al., 2024), BiasMedQA (Schmidgall et al., 2024), Hurtful Words (Zhang et al., 2020), Race-based Medicine (Omiye et al., 2023), Detect Debunked Stereotypes (Swaminathan et al., 2024), EquityMedQA (Pfohl et al., 2024), Superficial Fairness Alignment (Wei et al., 2024), Examines Biased AI (Adam et al.), Identify Biases (Yang et al., 2024b), Quantifying Cognitive Biases (Lin and Ng, 2023), Biases in Biomedical MLM (Kim et al., 2023), Bias of Disease Prediction (Zhao et al., 2024b), Racial and LGBTQ+ Biases (Xie et al., 2024), FairMedFM (Jin et al., 2024)
Explanability
Benchmarks
FaReBio (Fang et al., 2024), Pathway2Text (Yang et al., 2022)
Enhance Methods
Knowledge Graphs (Shariatmadari et al., 2024), Medical Imaging Explainability (Ghosh et al., 2023), MedExQA (Kim et al., 2024), Retrieval and Reasoning on KGs (Ji et al., 2024), DDCoT (Zheng et al., 2023), A ChatGPT Aided Explainable Framework (Liu et al., 2023a), Medical Concept-Driven Attention (Wang et al., 2022), FaReBio (Fang et al., 2024), LLM-GCE (He et al., 2024), kNN-Graph2Text (Yang et al., 2022), RAG-IM (Mahbub et al., 2024), MedThink (Gai et al., 2025)

Figure 2: Summary of the recent research across various dimensions of trustworthiness of LLMs in healthcare.

and Med-Gen, with Med-IE offering value in specific areas of trustworthiness.

2.3 Models

The models assessed in studies on trust in LLMs for the healthcare domain are outlined, along with their trustworthiness dimensions, in Appendix C, where we summarized key details such as the model name, release year, openness, architecture, task, and the institution responsible for its development. Figure 1(c) illustrates the proportions of different model types—open-source, closed-source, and architectures including encoder-only, decoder-only, and encoder-decoder—used in research addressing various trustworthiness aspects of LLMs in healthcare: Explainability, Fairness and Bias, Robustness, Safety, Privacy, and Truthfulness. From Figure 1(c), it is clear that Decoder-only and Open-source models are the most commonly used across all trustworthiness dimensions—especially in robustness, explainability, and truthfulness—highlighting their accessibility and alignment with generative tasks. Closed-source models appear more in fairness and privacy studies, while Encoder-only and Encoder-decoder models are used less frequently, mostly in fairness and truthfulness evaluations.

2.4 Tasks

The tasks covered various primary focuses of LLMs in healthcare. Inspired from the survey by Liu et al. (2024a), these tasks include:

Medical Information Extraction (Med-IE) Med-IE extracts structured medical data from unstructured sources such as EHRs, clinical notes, and research articles. Key tasks include entity recognition (identifying diseases, symptoms, and treatments), relationship extraction (understanding entity connections), event extraction (detecting clinical events and attributes), information summarization (condensing medical records), and adverse drug event detection (identifying medication-related risks).

Medical Question Answering (Med-QA) Med-QA systems interpret and respond to complex medical queries from patients, clinicians, and researchers. Their core functions include query understanding (interpreting user questions), information retrieval (finding relevant data in medical databases), and inference and reasoning (drawing conclusions, inferring relationships, and predicting outcomes based on retrieved data).

Medical Natural Language Inference (Med-NLI) Med-NLI analyzes the logical relationships between medical texts. Key tasks include textual entailment (determining if one statement logically follows another), contradiction detection (identifying conflicting statements), neutral relationship identification (recognizing unrelated statements), and causality recognition (inferring cause-and-effect relationships).

Medical Text Generation (Med-Gen) Med-Gen focuses on generating and summarizing medical content. Its key applications include text summarization (condensing lengthy documents into concise summaries) and content generation (producing new medical descriptions or knowledge based on input data).

3 Trustworthiness of LLMs in Healthcare

We examine the challenges related to the trustworthiness of LLMs in healthcare, outlining key strategies for identifying and mitigating these concerns. From our literature review screening, we identified truthfulness, privacy, safety, robustness, fairness and bias, and explainability as key trustworthiness dimensions of LLMs as highlighted in TrustLLM (Huang et al., 2024), particularly in healthcare. Figure 2 provides a summary of the recent research on trust in LLMs for healthcare across key dimensions of trustworthiness.

3.1 Truthfulness

Findings in Truthfulness

Current solutions like self-reflection and fact-checking frameworks reduce hallucinations, but they remain limited in scalability and generalizability. Existing methods lack robust grounding in long-form clinical contexts. Ensuring truthfulness will require hybrid verification pipelines that combine retrieval, reasoning, and multi-agent self-correction.

Ensuring the *truthfulness* of LLMs in healthcare is critical, as inaccurate or fabricated information can directly harm clinical decisions. Hallucinations typically emerge from biased data, poor contextual reasoning, or reliance on unverifiable sources (Ahmad et al., 2023), prompting the need for mechanisms that support factual correctness, source attribution, and uncertainty estimation. Recent efforts tackle these challenges through benchmarking, post-hoc correction, uncertainty quantification, and improved evidence synthesis—each targeting different aspects of factual reliability in medical LLMs.

Several benchmarks have emerged to quantify and categorize hallucinations. The Med-HALT benchmark (Pal et al., 2023) evaluates hallucination types using reasoning-based tests (e.g., "False Confidence") and memory checks. In multimodal settings, Med-HVL (Yan et al., 2024) distinguishes between Object Hallucination and Domain Knowledge Hallucination.

To mitigate hallucinations, post-hoc correction techniques are gaining traction. MEDAL (Li et al., 2024) presents a model-agnostic self-correction module that improves summarization outputs without retraining. Similarly, interactive feedback strategies like self-reflection loops (Ji et al., 2023) allow LLMs to iteratively refine their responses.

Uncertainty quantification approaches provide complementary detection tools. Farquhar et al. (2024) apply semantic entropy to flag low-confidence responses, while SEPs (Han et al., 2024a) offer a lightweight, hidden-state-based approximation suited for clinical use.

Recent efforts also examine the trustworthiness of evidence synthesis pipelines. Zhang et al. (2024) highlight risks when LLMs generate clinical summaries without grounding, emphasizing the need for transparency in literature retrieval and evidence aggregation. Debate-based evaluation, as explored in MAD (Smit et al., 2023), introduces multi-agent deliberation to vet factual consistency in medical QA. Finally, SEND (Mohammadzadeh et al., 2024) introduces a neuron dropout technique to detoxify hallucination-prone neurons during training, aiming to improve inherent model truthfulness.

Factual accuracy is critical for trust in healthcare LLMs, where clinical safety relies on reliable, verifiable outputs. Yet, current models often produce ungrounded content and lack source traceability. Recent work addresses this through medical claim benchmarks, self-correction, automated fact-checking, multi-turn verification, and multi-perspective reasoning—advancing transparency, factuality, and clinical relevance.

To support systematic validation, Akhtar et al. (2022) introduce PubHealthTab, a table-based dataset for checking public health claims against noisy evidence, while Sarrouti et al. (2021) propose

HEALTHVER, a benchmark for evidence-based fact-checking tailored to medical claims. These resources enable structured evaluation of LLM outputs and form the foundation for improving medical claim verification.

Beyond static benchmarks, dynamic self-correction methods have shown promise. Gou et al. (2024) propose CRITIC, a framework inspired by human fact-checking, in which LLMs iteratively assess and revise their own responses. This process mimics expert reasoning and introduces a layer of critical reflection into model outputs. Complementing this, Cohen et al. (2023) present a cross-examination approach, where a second "examiner" model engages in multi-turn dialogue to probe for factual inconsistencies in the original response. While CRITIC emphasizes human-like evaluation, cross-examination leverages interaction between models to simulate external verification.

To further reduce hallucinations and improve factual consistency, Tan et al. (2024) introduce a method that incorporates multiple scientific perspectives when resolving conflicting arguments, strengthening LLMs' reasoning capabilities through broader contextual understanding.

3.2 Privacy

Findings in Privacy

LLMs continue to pose serious privacy risks due to memorization and ineffective de-identification. While techniques like differential privacy and federated learning offer partial protection, they often degrade performance. Future solutions must enable fine-grained, instance-level privacy risk estimation across training and inference stages.

LLMs in healthcare pose significant *privacy* risks throughout their lifecycle, from pre-training to deployment, due to their tendency to memorize and potentially regenerate sensitive data such as protected health information (PHI) (Das et al., 2024; Pan et al., 2020). Key threats include data memorization, insufficient de-identification, and the privacy-utility trade-offs of fine-tuning methods. This section examines current vulnerabilities, mitigation strategies, and emerging approaches for achieving privacy-preserving healthcare LLMs.

Data memorization is a core concern, especially in domain-specific models like Medalpaca (Han et al., 2025), which are more likely to retain PHI and pose heightened re-identification risks (Yang et al., 2024a). Structured attacks like those demonstrated in SecureSQL (Song et al., 2024) reveal that even chain-of-thought (CoT) prompting provides only marginal defense against leakage.

Pre-training privacy measures include de-identification techniques like GPT-4 masking (Liu et al., 2023b) and synthetic note generation (Altalla' et al., 2025), though these offer limited protection. Xin et al. (2024) caution that such methods may create a false sense of security, as subtle semantic cues can still lead to PHI leakage.

Fine-tuning methods such as federated learning (Zhao et al., 2024a) and differential privacy (Singh et al., 2024) provide stronger safeguards by decentralizing data or adding noise to protect individual records. However, these methods often compromise model performance or scalability (Liu et al., 2024a).

Emerging techniques seek to reduce this trade-off. APNEAP (Wu et al., 2024) introduces activation patching for privacy neuron editing, reducing leakage without harming utility. Complementarily, Chen and Esmaeilzadeh (2024) offer a broader survey of privacy risks and solutions across generative AI use cases in healthcare.

Ethical and personalization challenges further complicate privacy design. Zhui et al. (2024) emphasize building privacy-conscious frameworks in medical education, while Kirk et al. (2024) caution that overly personalized alignment strategies may inadvertently violate user privacy, advocating instead for bounded personalization.

3.3 Safety

Findings in Safety

Medical LLMs can generate harmful or misleading content even after safety fine-tuning. Existing benchmarks reveal vulnerabilities to adversarial prompts and embedded misinformation. Ensuring safety requires proactive alignment strategies and multi-stage evaluation pipelines that simulate realistic clinical misuse scenarios.

Ensuring the *safety* of LLMs in healthcare is vital, as harmful outputs can lead to serious clinical consequences. Key concerns include the ease of injecting persistent falsehoods into model weights, inadequate performance on harmful prompts, trade-offs between safety alignment and hallucination, and privacy-related vulnerabilities that can escalate safety risks. This section explores current benchmarks, safety alignment strategies, and the overlap between safety and privacy threats. Han et al. (2024d) show that modifying just 1.1% of an LLM's weights can embed lasting biomedical falsehoods without affecting overall performance. Similarly, Han et al. (2024e) find that poisoning only 0.001% of training data can introduce persistent misinformation, highlighting the need for robust safeguards during training and deployment.

To systematically evaluate harmful outputs, benchmarks like MedSafetyBench (Han et al., 2024b) and Med-Harm (Han et al., 2024c) use adversarial and real-world queries to test model responses. Results show that even medically fine-tuned LLMs often fail safety criteria unless specifically optimized. MEDIC (Kanithi et al., 2024) broadens this evaluation across dimensions such as reasoning and reliability, offering a holistic safety diagnostic tool.

Safety alignment remains challenging due to its tension with other objectives. UNIWIZ (Das and Srihari, 2024) combines safety-driven training with fact-grounded retrieval to reduce unsafe outputs while maintaining factual accuracy. However, overalignment induces hallucinations, whereas underalignment allows unsafe behaviors, demonstrating the delicate balance required for clinical reliability.

Finally, privacy threats intersect with safety risks. Leemann et al. (2024) show that membership inference attacks, like Gradient Likelihood Ratio (GLiR), can detect whether individual patient data was used in training. This not only violates privacy but also raises safety concerns, as misuse of sensitive information can misguide clinical outcomes.

3.4 Robustness

Findings in Robustness

LLMs are fragile under distribution shifts, adversarial prompts, and instruction variations. Despite advances in adversarial testing and test-time adaptation, most defenses are brittle or task-specific. Achieving robustness demands context-aware evaluation, multi-agent training, and resilience to real-world perturbations.

Ensuring the *robustness* of LLMs is essential for their safe deployment in healthcare, where models must perform reliably across diverse clinical scenarios. Key challenges include adversarial vulnerability, sensitivity to domain shifts and instruction variations, and prompt-based attacks. To address these issues, recent work explores adversarial testing, test-time adaptation, prompt security, data augmentation, and instruction robustness strategies.

Adversarial robustness is addressed through synthetic data generation. Yuan et al. (2023) and Wang et al. (2024) introduce adversarial test samples tailored to the medical domain, such as synthetic anomaly cases and boundary stress testing, to assess model resilience. Alberts et al. (2023) emphasize the importance of aligning adversarial testing methods with real-world medical complexities. In parallel, Gallifant et al. (2024) reveal that simply substituting generic and brand drug names within biomedical benchmarks leads to performance drops of up to 10%, highlighting the fragility of LLMs to clinically trivial lexical shifts.

Uncertainty quantification offers another avenue for robustness. LLM-TTA (O'Brien et al., 2024) explores test-time adaptation techniques to enhance model performance on rare or unfamiliar cases, common in medical diagnostics. This approach complements adversarial robustness by identifying instances where models are likely to err.

Instruction robustness is examined by Ceballos-Arroyo et al. (2024), who find that specialized medical models may be more fragile than general-purpose models when instructions are reworded, suggesting that excessive domain adaptation may reduce flexibility.

Prompt security is enhanced by Tang et al. (2024), who introduce a framework that strengthens LLM robustness with cryptographic prompt authentication, mitigating vulnerabilities associated with prompt injections and adversarial attacks.

Data augmentation techniques are employed in MEDSAGE (Binici et al., 2025), which uses LLM-generated synthetic dialogues to simulate ASR errors, improving the robustness of medical dialogue summarization systems. Similarly, RobustQA (Han et al., 2023) benchmarks the robustness of domain adaptation for open-domain question answering across diverse domains, facilitating the evaluation of ODQA's domain robustness.

Lastly, prompt engineering strategies, such as out-of-context prompting, are explored by Cotta and Maddison (2024), who demonstrate that applying random counterfactual transformations can improve the fairness and robustness of LLM predictions without additional data or fine-tuning.

3.5 Fairness and Bias

Findings in Fairness

Biases in race, gender, and identity persist across medical LLM outputs. While new benchmarks and mitigation methods help, many remain narrow in scope or poorly aligned with clinical realities. Fairness must be pursued through intersectional audits, inclusive datasets, and collaboration with impacted communities.

Ensuring *fairness* in LLMs is critical for equitable healthcare, as biased predictions can reinforce existing disparities in access, diagnosis, and treatment. Key areas of concern include demographic bias (e.g., race, gender, identity), automated detection of these biases, mitigation strategies based on model accessibility, and the need for ethical clarity and conceptual frameworks. Recent work spans benchmark creation, debiasing techniques, prompt interventions, and calls for more transparent fairness evaluations.

Bias identification remains a foundational step. Studies show that LLMs can replicate and even amplify racial, gender, and identity-based biases. For example, Omiye et al. (2023), Zack et al. (2024), and Kim et al. (2023) highlight persistent demographic biases in medical responses. Zhao et al. (2024b) find that diagnostic recommendations vary unfairly by demographic group, while Xie et al. (2024) reveal systematic inequities in outputs concerning race and LGBTQ+ identities. Patel et al. (2024) further demonstrate that LLMs can reinforce social and gender-based stereotypes in sensitive areas such as sexual and reproductive health, underscoring the risks in context-dependent medical interactions.

Detection and benchmarking tools help quantify and monitor these disparities. Swaminathan et al. (2024) propose tools for identifying race-based stereotypes in medical Q&A. Benchmarks such as BiasMedQA (Schmidgall et al., 2024), EquityMedQA (Pfohl et al., 2024), and FairMedFM (Jin et al., 2024) offer frameworks for testing model behavior across diverse patient profiles and clinical contexts.

Mitigation strategies differ by model accessibility. For open-source models, techniques like adversarial debiasing (Zhang et al., 2020), causal intervention (CI4MRC) (Zhu et al., 2023), multi-agent collaboration (Ke et al., 2024), and attribute-neutral modeling (Hu et al., 2024) are applied to reduce bias. Data augmentation (Parray et al., 2023) and bias-aware embedding assessments (Lin and Ng, 2023) provide further tools to enhance fairness in pretraining and inference.

Closed-source models present unique challenges due to limited transparency. In these cases, fairness is addressed via instruction fine-tuning (Singhal et al., 2023), external prompt engineering (Schmidgall et al., 2024), or bounded personalization strategies (Kirk et al., 2024), though these are less interpretable and harder to audit.

Ethical and conceptual considerations also play a role. Wei et al. (2024) call for distinguishing between intrinsic and behavioral fairness, while Zhui et al. (2024) and Cotta and Maddison (2024) promote fairness through education and prompt design. Finally, Adam et al. and Yang et al. (2024b) warn that unchecked bias can distort care decisions and patient trust, emphasizing the stakes of fairness in real-world applications.

3.6 Explanability

Findings in Explainability

Despite progress in rationale generation and attention visualization, most explainability tools lack clinical relevance or faithfulness. Current methods often fail to align with clinician reasoning. Future work must bridge this gap with domain-specific frameworks and causal or counterfactual explanation techniques.

The lack of *explainability* in LLMs hinders clinical adoption by limiting transparency and trust. Recent research explores both intrinsic (model-integrated) and post-hoc (output-interpretation) techniques to make LLM reasoning more interpretable. These methods span a wide range of modalities including text, graphs, tables, and images—and often incorporate domain-specific knowledge or human-centered reasoning to bridge model outputs and clinical expectations.

Intrinsic explainability methods enhance transparency by aligning model attention with medical knowledge. For example, Shariatmadari et al. (2024) integrate knowledge graphs with attention visualization, while Wang et al. (2022) use Wikipedia-derived medical concepts to guide attention for code prediction, resulting in more concept-consistent outputs. Similarly, structure-to-text models like Pathway2Text (Yang et al., 2022) convert biomedical graphs into interpretable narratives, supporting more intuitive understanding of complex structured inputs.

Post-hoc strategies focus on generating faithful rationales and justifications. FaReBio (Fang et al., 2024) highlights how summarization faithfulness suffers with increased abstractiveness and introduces a benchmark to evaluate reasoning fidelity. In the molecular domain, LLM-GCE (He et al., 2024) generates counterfactuals for Graph Neural Networks (GNNs) using dynamic feedback to ensure chemically valid, interpretable explanations.

Several methods target zero-shot interpretability without task-specific fine-tuning. RAG-IM (Mahbub et al., 2024) enables table-based clinical predictions with natural language justifications, while Liu et al. (2023a) embed ChatGPT into a diagnostic workflow with integrated interpretability components. Retrieval-based systems such as Retrieval + KG (Ji et al., 2024) and DDCoT (Zheng et al., 2023) further enhance reasoning by chaining knowledge-grounded prompts across modalities.

Explainability in imaging and multimodal contexts is also gaining traction. MedThink (Gai et al., 2025) fuses visual and textual inputs to improve multimodal reasoning, and MedExQA (Kim et al., 2024) supplies detailed rationales for visual question answering. Ghosh et al. (2023) decomposes black-box decisions into expert modules with first-order logic (FOL) reasoning.

4 Future Directions

While core trust dimensions—truthfulness, privacy, robustness, fairness, explainability, and safety—have been the focus of recent work, emerging model paradigms such as multi-agent systems, multi-modal models, and small open-source LLMs introduce new trust challenges underexplored.

Multi-Agent LLMs Multi-agent LLMs enable distributed reasoning through collaboration between specialized agents, offering improved robustness and self-correction. However, they also raise concerns around coordination, error propagation, and the interpretability of inter-agent communication. Trustworthy multi-agent systems will require protocols for communication, verification, and evaluation that ensure factual alignment and fairness. For example, Lu et al. (2024) introduce TriageAgent, a clinical multi-agent framework with role-specific LLMs for diagnosis and decision-making. While it shows benefits like structured collaboration and early stopping, it also reveals trust challenges including inconsistent agent confidence, limited transparency, and error propagation—highlighting the need for stronger verification and alignment in high-stakes settings.

Multimodal Foundation Models Multi-modal LLMs combine text, images, and structured data, better reflecting real-world clinical inputs but complicating trust evaluation. Challenges include cross-modal hallucination, misalignment, and reduced explainability. Addressing these issues will require modality-specific assessments, interpretable fusion strategies, and fairness testing across both textual and visual modalities. For example, Liu et al. (2024b) evaluate open-source multimodal LLMs for genomics and proteomics, highlighting issues with factual consistency and alignment across modalities—underscoring the importance of structured evaluation and interpretable model design in biomedical contexts.

Small Open-Source LLMs Small open-source medical LLMs are gaining traction for their transparency, adaptability, and lower computational demands, making them attractive for deployment in resource-constrained or privacy-sensitive settings. However, their reduced capacity often leads to increased hallucinations, weaker safety alignment, and heightened privacy risks during fine-tuning on limited clinical data. Ensuring their trustworthiness requires lightweight hallucination mitigation, privacy-preserving training, and scalable evaluation pipelines. Despite their growing use, few studies directly examine these trust issues in small medical LLMs, as most existing research focuses on larger or general-purpose models—leaving a critical gap in the literature.

5 Conclusion

As large language models continue to expand their role in healthcare, ensuring their trustworthiness remains a critical challenge. This survey reviewed six core dimensions—truthfulness, privacy, safety, robustness, fairness, and explainability—highlighting key methods, benchmarks, and limitations in current research. While recent advances have laid important groundwork, most existing solutions remain narrowly scoped and lack integration across dimensions, limiting their effectiveness in real-world clinical settings.

Limitations

This survey provides a comprehensive overview of the challenges associated with LLMs in healthcare, but it primarily focuses on existing methodologies, leaving out emerging technologies that could address these issues in new ways. It also lacks practical insights into the real-world implementation of these solutions, such as deployment challenges, cost considerations, and system integration, which would make the findings more applicable to healthcare settings.

While the paper addresses privacy and safety, it does not fully explore broader ethical issues like informed consent, patient autonomy, and human oversight. Additionally, the survey focuses on current research without delving into the long-term societal and health impacts of LLM deployment, such as changes in doctor-patient relationships, patient trust, and healthcare workflows.

References

Hammaad Adam, Aparna Balagopalan, Emily Alsentzer, Fotini Christia, and Marzyeh Ghassemi. Just following ai orders: When unbiased people are influenced by biased ai. In *Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022*.

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

Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. 2022. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States. Association for Computational Linguistics.

Ian L Alberts, Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi, Axel Rominger, and Ali Afshar-Oromieh. 2023. Large language models (llm) and chatgpt: what will the impact on nuclear medicine be? *European journal of nuclear medicine and molecular imaging*, 50(6):1549–1552.

Bayan Altalla', Sameera Abdalla, Ahmad Altamimi, Layla Bitar, Amal Al Omari, Ramiz Kardan, and Iyad Sultan. 2025. Evaluating gpt models for clinical note de-identification. *Scientific Reports*, 15(1):3852.

Zhenyu Bi, Sajib Acharjee Dip, Daniel Hajialigol, Sindhura Kommu, Hanwen Liu, Meng Lu, and Xuan Wang. 2024. Ai for biomedicine in the era of large language models. *arXiv preprint arXiv:2403.15673*.

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T. Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F. Chen, and Stefan Winkler. 2025. MEDSAGE: Enhancing robustness of medical dialogue summarization to ASR errors with LLM-generated synthetic dialogues. In *AI4X 2025 International Conference*.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered McInerney, Byron C. Wallace, and Silvio Amir. 2024. Open (clinical) LLMs are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71, Bangkok, Thailand. Association for Computational Linguistics.

Yan Chen and Pouyan Esmaeilzadeh. 2024. Generative ai in medical practice: in-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26:e53008.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross examination. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Leonardo Cotta and Chris J Maddison. 2024. Out-of-context prompting boosts fairness and robustness in large language model predictions. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Badhan Chandra Das, M Hadi Amini, and Yanzhao Wu. 2024. Security and privacy challenges of large language models: A survey. *arXiv preprint arXiv:2402.00888*.

Souvik Das and Rohini K Srihari. 2024. Uniwiz: A unified large language model orchestrated wizard for safe knowledge grounded conversations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1749–1762.

Biaoyan Fang, Xiang Dai, and Sarvnaz Karimi. 2024. Understanding faithfulness and reasoning of large language models on plain biomedical summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9890–9911, Miami, Florida, USA. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. Medthink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7438–7450.

Jack Gallifant, Shan Chen, Pedro José Ferreira Moreira, Nikolaj Munch, Mingye Gao, Jackson Pond, Leo Anthony Celi, Hugo Aerts, Thomas Hartvigsen, and Danielle Bitterman. 2024. Language models are surprisingly fragile to drug names in biomedical benchmarks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12448–12465, Miami, Florida, USA. Association for Computational Linguistics.

Shantanu Ghosh, Ke Yu, Forough Arabshahi, and kayhan Batmanghelich. 2023. Bridging the gap: From post hoc explanations to inherently interpretable models for medical imaging. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. 2024a. Semantic entropy probes: Robust and cheap hallucination detection in llms. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. RobustQA: Benchmarking the robustness of domain adaptation for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4294–4311, Toronto, Canada. Association for Computational Linguistics.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024b. Medsafetybench: Evaluating and improving the medical safety of large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tessa Han, Aounon Kumar, Chirag Agarwal, and Himabindu Lakkaraju. 2024c. Towards safe large language models for medicine. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figueroa, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2025. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

Tianyu Han, Sven Nebelung, Firas Khader, Tianci Wang, Gustav Müller-Franzes, Christiane Kuhl, Sebastian Försch, Jens Kleesiek, Christoph Haarburger, Keno K Bressem, et al. 2024d. Medical large language models are susceptible to targeted misinformation attacks. *NPJ Digital Medicine*, 7(1):288.

Xiang Han, Qi Zhang, Kai Wang, Yitong Zhang, Chenyu Guo, Dongdong Chen, Xinyang Liu, and James Zou. 2024e. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*.

Yinhan He, Zaiyi Zheng, Patrick Soga, Yaochen Zhu, Yushun Dong, and Jundong Li. 2024. Explaining graph neural networks with large language models: A counterfactual perspective on molecule graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7079–7096, Miami, Florida, USA. Association for Computational Linguistics.

Lianting Hu, Dantong Li, Huazhang Liu, Xuanhui Chen, Yunfei Gao, Shuai Huang, Xiaoting Peng, Xueli Zhang, Xiaohe Bai, Huan Yang, et al. 2024. Enhancing fairness in ai-enabled medical systems with the attribute neutral framework. *Nature Communications*, 15(1):8767.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Position: Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270. PMLR.

Yixin Ji, Kaixin Wu, Juntao Li, Wei Chen, Mingjie Zhong, Xu Jia, and Min Zhang. 2024. Retrieval and reasoning on KGs: Integrate knowledge graphs into large language models for complex question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7598–7610, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. 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.

Ruinan Jin, Zikang Xu, Yuan Zhong, Qingsong Yao, Qi Dou, S Kevin Zhou, and Xiaoxiao Li. 2024. FairmedFM: Fairness benchmarking for medical imaging foundation models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. *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.

Michelle Kim, Junghwan Kim, and Kristen Johnson. 2023. Race, gender, and age biases in biomedical masked language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11806–11815, Toronto, Canada. Association for Computational Linguistics.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. MedExQA: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 167–181, Bangkok, Thailand. Association for Computational Linguistics.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392.

Tobias Leemann, Bardh Prenkaj, and Gjergji Kasneci. 2024. Is my data safe? predicting instance-level membership inference success for white-box and black-box attacks. In *ICML 2024 Next Generation of AI Safety Workshop*.

Songda Li, Yunqi Zhang, Chunyuan Deng, Yake Niu, and Hui Zhao. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Ruixi Lin and Hwee Tou Ng. 2023. Mind the biases: Quantifying cognitive biases in language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5269–5281, Toronto, Canada. Association for Computational Linguistics.

Jiaxiang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, YANG FENG, and Zuozhu Liu. 2023a. A chatGPT aided explainable framework for zero-shot medical image diagnosis. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*.

Lei Liu, Xiaoyan Yang, Junchi Lei, Xiaoyang Liu, Yue Shen, Zhiqiang Zhang, Peng Wei, Jinjie Gu, Zhixuan Chu, Zhan Qin, et al. 2024a. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

Tianyu Liu, Yijia Xiao, Xiao Luo, Hua Xu, Wenjin Zheng, and Hongyu Zhao. 2024b. Geneverse: A collection of open-source multimodal large language models for genomic and proteomic research. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4819–4836, Miami, Florida, USA. Association for Computational Linguistics.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. 2023b. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Meng Lu, Brandon Ho, Dennis Ren, and Xuan Wang. 2024. TriageAgent: Towards better multi-agents collaborations for large language model-based clinical triage. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5747–5764, Miami, Florida, USA. Association for Computational Linguistics.

Sazan Mahbub, Caleb Ellington, Sina Alinejad, Kevin Wen, Yingtao Luo, Ben Lengerich, and Eric P. Xing. 2024. From one to zero: RAG-IM adapts language models for interpretable zero-shot predictions on clinical tabular data. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

Shahrad Mohammadzadeh, Juan David Guerra, Marco Bonizzato, Reihaneh Rabbany, and Golnoosh Farnadi. 2024. Hallucination detox: Sensitive neuron dropout (send) for large language model training. In *Neurips Safe Generative AI Workshop 2024*.

Kyle O'Brien, Nathan Ng, Isha Puri, Jorge Mendez, Hamid Palangi, Yoon Kim, Marzyeh Ghassemi, and Thomas Hartvigsen. 2024. Improving black-box robustness with in-context rewriting. *arXiv preprint arXiv:2402.08225*.

Jesutofunmi A Omiye, Jenna C Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6(1):195.

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, Singapore. Association for Computational Linguistics.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Ateeb Ahmad Parray, Zuhrat Mahfuza Inam, Diego Ramonfaur, Shams Shabab Haider, Sabuj Kanti Mistry, and Apurva Kumar Pandya. 2023. Chatgpt and global public health: applications, challenges, ethical considerations and mitigation strategies.

Parth Patel, Nafise Moosavi, and Leon Derczynski. 2024. Evaluating biases in context-dependent sexual and reproductive health questions. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12):3590–3600.

Maxx Richard Rahman, Ruoxuan Liu, and Wolfgang Maass. 2024. Incorporating metabolic information into LLMs for anomaly detection in clinical time-series. In *NeurIPS Workshop on Time Series in the Age of Large Models*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Schmidgall, Carl Harris, Ime Essien, Daniel Olshvang, Tawsifur Rahman, Ji Woong Kim, Rojin Ziaei, Jason Eshraghian, Peter Abadir, and Rama Chellappa. 2024. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1):295.

Amir Hassan Shariatmadari, Sikun Guo, Sneha Srinivasan, and Aidong Zhang. 2024. Harnessing the power of knowledge graphs to enhance llm explainability in the biomedical domain,(2024). *IJACSA) International Journal of Advanced Computer Science and Applications*.

Tanmay Singh, Harshvardhan Aditya, Vijay K Madisetti, and Arshdeep Bahga. 2024. Whispered tuning: Data privacy preservation in fine-tuning llms through differential privacy. *Journal of Software Engineering and Applications*, 17(1):1–22.

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.

Andries Petrus Smit, Paul Duckworth, Nathan Grinsztajn, Kale-ab Tessera, Thomas D Barrett, and Arnu Pretorius. 2023. Are we going mad? benchmarking multi-agent debate between language models for medical q&a. In *Deep Generative Models for Health Workshop NeurIPS 2023*.

Yanqi Song, Ruiheng Liu, Shu Chen, Qianhao Ren, Yu Zhang, and Yongqi Yu. 2024. SecureSQL: Evaluating data leakage of large language models as natural language interfaces to databases. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5975–5990, Miami, Florida, USA. Association for Computational Linguistics.

Akshay Swaminathan, Sid Salvi, Philip Chung, Alison Callahan, Suhana Bedi, Alyssa Unell, Mehr Kashyap, Roxana Daneshjou, Nigam Shah, and Dev Dash. 2024. Feasibility of automatically detecting practice of race-based medicine by large language models. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Neşet Özkan Tan, Niket Tandon, David Wadden, Oyvind Tafjord, Mark Gahegan, and Michael Witbrock. 2024. Faithful reasoning over scientific claims. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 263–272.

Ruixiang Tang, Yu-Neng Chuang, Xuanting Cai, Mengnan Du, and Xia Hu. 2024. Secure your model: An effective key prompt protection mechanism for large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4061–4073, Mexico City, Mexico. Association for Computational Linguistics.

Tao Wang, Linhai Zhang, Chenchen Ye, Junxi Liu, and Deyu Zhou. 2022. A novel framework based on medical concept driven attention for explainable medical code prediction via external knowledge. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1407–1416, Dublin, Ireland. Association for Computational Linguistics.

Yichen Wang, Shangbin Feng, Abe Bohan Hou, Xiao Pu, Chao Shen, Xiaoming Liu, Yulia Tsvetkov, and Tianxing He. 2024. Stumbling blocks: Stress testing the robustness of machine-generated text detectors under attacks. *arXiv preprint arXiv:2402.11638*.

Qiyao Wei, Alex James Chan, Lea Goetz, David Watson, and Mihaela van der Schaar. 2024. Actions speak louder than words: Superficial fairness alignment in LLMs. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Xinwei Wu, Weilong Dong, Shaoyang Xu, and Deyi Xiong. 2024. Mitigating privacy seesaw in large language models: Augmented privacy neuron editing via activation patching. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5319–5332, Bangkok, Thailand. Association for Computational Linguistics.

Sean Xie, Saeed Hassanpour, and Soroush Vosoughi. 2024. Addressing healthcare-related racial and LGBTQ+ biases in pretrained language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4451–4464, Mexico City, Mexico. Association for Computational Linguistics.

Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2024. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. In *Neurips Safe Generative AI Workshop 2024*.

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

Junwei Yang, Zequn Liu, Ming Zhang, and Sheng Wang. 2022. Pathway2Text: Dataset and method for biomedical pathway description generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1441–1454, Seattle, United States. Association for Computational Linguistics.

Xinyu Yang, Zichen Wen, Wenjie Qu, Zhaorun Chen, Zhiying Xiang, Beidi Chen, and Huaxiu Yao. 2024a.

Memorization and privacy risks in domain-specific large language models. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024b. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, 4(1).

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Gongbo Zhang, Qiao Jin, Denis Jered McInerney, Yong Chen, Fei Wang, Curtis L Cole, Qian Yang, Yanshan Wang, Bradley A Malin, Mor Peleg, et al. 2024. Leveraging generative ai for clinical evidence synthesis needs to ensure trustworthiness. *Journal of Biomedical Informatics*, 153:104640.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

Jujia Zhao, Wenjie Wang, Chen Xu, Zhaochun Ren, See-Kiong Ng, and Tat-Seng Chua. 2024a. Llm-based federated recommendation. *arXiv preprint arXiv:2402.09959*.

Yutian Zhao, Huimin Wang, Yuqi Liu, Wu Suhuang, Xian Wu, and Yefeng Zheng. 2024b. Can LLMs replace clinical doctors? exploring bias in disease diagnosis by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13914–13935, Miami, Florida, USA. Association for Computational Linguistics.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

Jiazheng Zhu, Shaojuan Wu, Xiaowang Zhang, Yuexian Hou, and Zhiyong Feng. 2023. Causal intervention for mitigating name bias in machine reading comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12837–12852.

Li Zhui, Li Fenghe, Wang Xuehu, Fu Qining, and Ren Wei. 2024. Ethical considerations and fundamental principles of large language models in medical education. *Journal of Medical Internet Research*, 26:e60083.

A Inclusion & Exclusion Criteria Details

We conducted an extensive search to identify peer-reviewed papers that address the trustworthiness of LLMs in the healthcare domain. Our search strategy involved a wide range of keyword combinations related to LLMs and core trust dimensions, including: trustworthiness, truthfulness, privacy, safety, robustness, fairness, bias, and explainability. We targeted both domain-specific and general AI venues, focusing on recent publications from 2022 onward.

Specifically, we searched across top-tier conferences and journals, including ACL, EMNLP, NAACL, ICML, NeurIPS, ICLR, KDD, AAAI, IJCAI, Nature, and Science, using platforms such as Google Scholar, Nature, and Science. A full list of keyword queries used in our search is provided below. These queries combined domain terms (medical, clinical) with trust-related dimensions, applied across both "large language models" and "foundation models." Examples include:

- large language models, medical, explainability
- large language models, medical, explainable
- foundation model, medical, explainability
- large language models, clinical, explainability
- large language models, medical, truthfulness
- large language models, medical, trustworthiness
- foundation model, medical, trustworthiness
- large language models, clinical, truthfulness
- large language models, clinical, safety
- large language models, medical, safety
- foundation model, medical, safety
- large language models, clinical, fairness
- large language models, medical, fairness
- foundation model, medical, fairness
- large language models, clinical, robustness
- foundation model, medical, robustness
- large language models, medical, robustness
- large language models, clinical, privacy
- large language models, medical, privacy
- foundation model, medical, privacy
- large language models, clinical, ethics
- large language models, medical, ethics
- foundation model, medical, ethics

In total, our initial search returned approximately 15,322 results, including duplicates and non-relevant papers. Our filtering process proceeded in three stages:

- Duplicate removal – approximately 11,172 papers eliminated.
- Relevance screening – we excluded papers that: (a) did not focus on trustworthiness aspects (e.g., architecture design or multi-modal fusion techniques), (b) were not specific to the healthcare domain, or (c) were unpublished preprints (e.g., arXiv manuscripts).
- Final selection – we curated a final set of 62 papers that directly addressed trust-related challenges in healthcare LLMs, focusing on one or more of the following dimensions: truthfulness, privacy, safety, robustness, fairness, bias, and explainability.

B Comparison of Datasets

We systematically collected and analyzed 38 datasets relevant to the study of trust in LLMs for healthcare. Table 1 provides a comprehensive summary, highlighting key attributes such as data type, content, associated tasks, and the specific trustworthiness dimensions they address. These datasets vary widely, including web-scraped data, curated domain-specific datasets, public text corpora, synthetic data, real-world data, and private datasets. Each dataset's content specifies its composition, while its associated task defines its primary research application. Additionally, we categorize the datasets based on critical trustworthiness dimensions—truthfulness, privacy and safety, robustness, fairness and bias, and explainability—offering a structured evaluation of their contributions to building reliable and trustworthy healthcare AI.

Datasets	Data Type	Content	Task	Dimensions
MultiMedQA	Combination of Public and Synthetic Data, Curated Domain-Specific Dataset	208,000 entries. A benchmark combining six existing medical questions answering datasets spanning professional medicine, research and consumer queries and a new dataset of medical questions searched online, HealthSearchQA.	(Med-QA) Tasks including Medical Question Answering, Clinical Reasoning, Evidence-Based Medicine, Multilingual and Multimodal Support, Bias and Safety Analysis	Fairness and Bias
BiasMedQA	Curated Domain-Specific Datasets	1273 USMLE questions	(Med-QA) Replicate common clinically relevant cognitive biases	Fairness and Bias
EquityMedQA	Curated domain-specific datasets and synthetic data	4,619 examples. Cover a wide range of medical topics to surface biases that could harm health equity, including implicit and explicit adversarial questions addressing biases like stereotypes, lack of structural explanations, and withholding information.	(Med-QA) Evaluate the performance of LLMs in generating unbiased, equitable medical responses.	Fairness and Bias
SQuAD	Curated Domain-Specific Dataset	Consists of over 100,000 question-answer pairs derived from more than 500 articles from Wikipedia. Each question is paired with a segment of text from the corresponding article, serving as the answer.	(Med-QA)To develop models that can read a passage and answer questions about it, assessing the model's ability to understand and extract information from the text.	Fairness and Bias
MIMIC- III	Public text corpora, real-world data	De-identified health-related data from over 40,000 critical care patients, including demographics, vital signs, laboratory tests, medications, and caregiver notes.	(Med-IE) Epidemiological studies, clinical decision-rule improvement, machine learning in healthcare.	Fairness and Bias, Explainability, Robustness
MedQA	Curated Domain-Specific Datasets	194,000 multiple-choice medical exam questions. A benchmark that includes questions drawn from the United States Medical License Exam (USMLE).	(Med-QA) Exam the physicians to test their ability to make clinical decisions	Fairness and Bias, Robustness, Explainability, Truthfulness, Privacy
PMC-Patients	Curated dataset derived from public text corpora.	Contains 167,000 patient summaries extracted from 141,000 PMC articles	(Med-IE) Designed to benchmark ReCDS systems through two primary tasks: Patient-to-Article Retrieval (PAR), Patient-to-Patient Retrieval (PPR)	Robustness
MedSafetyBench	Curated domain-specific dataset and synthetic (generated using GPT-4, Llama-2-7b-chat, and adversarial techniques).	1,800 harmful medical requests violating medical ethics, along with 900 corresponding safe responses. The dataset is structured based on the Principles of Medical Ethics from the American Medical Association (AMA).	(Med-Gen) Assess the medical safety of LLMs by testing whether they refuse to comply with harmful medical requests. Fine-tune LLMs using medical safety demonstrations to enhance their alignment with ethical medical guidelines.	Safety

Datasets	Data Type	Content	Task	Dimensions
UNIWIZ	Synthetic and curated data, including: 17,638 quality-controlled conversations, and 10,000 augmented preference data	17,638 conversations and 10,000 augmented preference data. Features conversations that integrate safety and knowledge alignment. A "safety-priming" method was employed to generate synthetic safety data, and factual information was injected into conversations by retrieving content from curated sources.	(Med-Gen) Fine-tune large language models to enhance their performance in generating safe and knowledge-grounded conversations.	Safety
SciFact	Curated Domain-Specific Dataset.	2,011 claims. Includes claims and corresponding evidence abstracts, each annotated with labels indicating whether the claim is supported or refuted, along with rationales justifying the decision.	(Med-Gen) To verify the veracity of scientific claims by identifying supporting or refuting evidence within abstracts and providing justifications for these decisions.	Truthfulness
PubHealthTab	Curated Domain-Specific Dataset	Contains 1,942 real-world public health claims, each paired with evidence tables extracted from over 300 websites.	(Med-Gen) Facilitates evidence-based fact-checking by providing claims and corresponding evidence tables for verification.	Truthfulness
LAMA	Curated Domain-Specific Dataset.	24,223 entries of knowledge sources. Comprises a set of knowledge sources, each containing a collection of facts.	(Med-Gen) To probe pre-trained language models to determine the extent of their factual and commonsense knowledge.	Truthfulness
TriviaQA	Curated Domain-Specific Dataset.	Consists of over 650,000 question-answer pairs, each linked to a set of supporting documents. The questions are sourced from trivia websites, and the answers are derived from the corresponding documents.	(Med-QA) Training and evaluating models on reading comprehension, specifically focusing on the ability to extract and reason over information from provided documents to answer questions.	Truthfulness
Natural Questions (NQ)	Real data	99.80 GB, with downloaded files accounting for 45.07 GB and the generated dataset occupying 54.73 GB. consists of real anonymized queries from Google's search engine users, paired with answers derived from entire Wikipedia articles.	(Med-QA) To develop and evaluate question-answering systems that can read and comprehend entire Wikipedia articles to find answers to user queries.	Truthfulness
PopQA	Curated Domain-Specific Dataset.	consists of 14,000 QA pairs, each associated with fine-grained Wikidata entity IDs, Wikipedia page views, and relationship type information.	(Med-QA) Designed for open-domain question answering tasks, focusing on evaluating the effectiveness of language models in retrieving and utilizing factual knowledge.	Truthfulness
FEVER	Curated Domain-Specific Dataset.	comprises 185,000 claims, each paired with evidence from Wikipedia articles. These claims are categorized as supported, refuted, or not verifiable.	(Med-Gen) Fact extraction and verification, where models are trained to determine the veracity of claims based on provided evidence.	Truthfulness

Datasets	Data Type	Content	Task	Dimensions
HEALTHVER	Curated Domain-Specific Dataset.	contains 14,330 evidence-claim pairs labeled as SUPPORTS, REFUTES, or NEUTRAL, derived from real-world health claims, mainly about COVID-19, verified against scientific articles.	(Med-Gen) Training and evaluating models on the task of verifying the truthfulness of health-related claims by assessing their alignment with scientific evidence. This involves classifying claims as supported, refuted, or neutral based on the provided evidence.	Truthfulness
Med-HALT	Synthetic and Real Data, Curated Domain-Specific Dataset, and Public Dataset	59,254 entries. Consist of Reasoning-Based Assessments, Memory-Based Assessments, Medical Scenarios, Evaluation Metrics	(Med-Gen) Tasks including Evaluation of Hallucination in Medical AI, Reliability Benchmarking, Error Analysis, Mitigation Development	Truthfulness
MedICaT	Public Text Corpora And Real Data (curated from publicly available biomedical literature)	217,060 figures extracted from 131,410 open-access papers. Contains medical images (e.g., radiographs, charts, and diagrams) paired with captions extracted from biomedical literature. Also, includes metadata about the source and context of the images.	(Med-Gen) Task including Medical Image Captioning, Text-Image Retrieval, Medical Reasoning	Truthfulness
BioASQ	Curated Domain-Specific Dataset; Real Data.	3,743 training questions and 500 test questions. The dataset comprises English-language biomedical questions, each accompanied by reference answers and related materials. These questions are designed to reflect real information needs of biomedical experts, making the dataset both realistic and challenging.	(Med-QA) The primary task is Biomedical Question Answering (QA), which involves systems providing accurate answers to questions based on biomedical data. The dataset supports various QA tasks, including yes/no, factoid, list, and summary questions.	Truthfulness
FactualBio	Synthetic Data; Public Text Corpora.	collection of biographies of individuals notable enough to have Wikipedia pages but lacking extensive detailed coverage. The dataset was generated using GPT-4 and includes biographies of 21 individuals randomly sampled from the WikiBio dataset.	(Med-Gen) Evaluating the factual accuracy of language models, particularly in the context of biography generation. It serves as a benchmark for detecting hallucinations and assessing the factual consistency of generated text.	Truthfulness
PubMedQA	Curated Domain-Specific Dataset.	Consists of over 1,000 question-answer pairs derived from PubMed abstracts, focusing on various biomedical topics.	(Med-QA) Evaluates the ability of models to comprehend and extract information from biomedical texts to answer specific questions.	Truthfulness
MedQuAD	Curated Domain-Specific Dataset.	The dataset encompasses 37 question types, such as Treatment, Diagnosis, and Side Effects, associated with diseases, drugs, and other medical entities like tests.	(Med-QA) Designed for medical question answering, the dataset aids in developing and evaluating systems that can understand and respond to medical inquiries.	Truthfulness
LiveMedQA2017	Curated Domain-Specific Dataset	Consists of 634 question-answer pairs corresponding to National Library of Medicine (NLM) questions	(Med-QA) Medical question answering, focusing on consumer health questions received by the U.S. National Library of Medicine.	Truthfulness

Datasets	Data Type	Content	Task	Dimensions
MASH-QA	Curated Domain-Specific Dataset.	Approximately 25,000 question-answer pairs sourced from WebMD, covering a wide range of healthcare topics.	(Med-QA) Designed for multiple-answer span extraction in healthcare question answering.	Truthfulness
SecureSQL	Curated domain-specific dataset	Comprises meticulously annotated samples, including both positive and negative instances. The dataset encompasses 57 databases across 34 diverse domains, each associated with specific security conditions.	(Med-IE) Evaluate and analyze data leakage risks in LLMs, particularly concerning SQL query generation and execution.	Privacy
Medical Meadow	curated domain-specific dataset	It comprises approximately 1.5 million data points across various tasks, including question-answer pairs generated from openly available medical data using models like OpenAI's	(Med-Gen) Designed to enhance large language models (LLMs) for medical applications	Privacy
Electronic Health Records (EHR) at (KHCC)	Private dataset	gpt-3.5-turbo	(Med-IE) Clinical research, outcome analysis.	Privacy
MedVQA	Curated domain-specific dataset	794 image-question-answer triplets. A collection of medical visual question answering pairs, designed to train and evaluate models that interpret medical images and answer related questions.	(Med-QA) Visual question answering, medical image understanding.	Explainability
MedExQA	Curated domain-specific dataset	965 multiple-choice medical questions. A dataset focused on medical examination questions and answers, intended to aid in the development of AI models for medical exam preparation and assessment.	(Med-QA) Question answering, educational assessment.	Explainability
MedMCQA	Curated domain-specific dataset	194,000 multiple-choice questions from AIIMS and NEET PG entrance exams, covering 2,400 healthcare topics across 21 medical subjects. A multiple-choice question-answering dataset in the medical domain, aimed at training models to handle medical examinations and practice questions.	(Med-QA) Multiple-choice question answering, medical education.	Explainability
TCM Medical Licensing Examination(MLE)	Curated domain-specific dataset	600 multiple-choice questions. A dataset comprising questions and answers from Traditional Chinese Medicine licensing examinations.	(Med-QA) Educational assessment, question answering.	Explainability

Datasets	Data Type	Content	Task	Dimensions
Pneumonia Dataset	Curated domain-specific dataset	5,863 images. Medical images (such as chest X-rays) labeled for the presence or absence of pneumonia, used for training diagnostic models.	(Med-IE) Image classification, disease detection.	Explainability
Montgomery Dataset	Curated domain-specific dataset	X-ray Set comprises 138 posterior-anterior chest X-ray images, with 80 normal and 58 abnormal cases indicative of tuberculosis. Chest X-ray images with manual segmentations of the lung fields, useful for pulmonary research.	(Med-IE) Image segmentation, tuberculosis detection.	Explainability
Shenzhen Dataset	Curated domain-specific dataset	Chest X-ray dataset comprises 662 frontal chest X-rays, including 326 normal cases and 336 cases with manifestations of tuberculosis. Chest X-ray images collected in Shenzhen, China, with annotations for tuberculosis manifestations.	(Med-IE) Disease classification, image analysis.	Explainability
IDRID Dataset	Curated domain-specific dataset	1,113 images. Retinal images with annotations for diabetic retinopathy lesions, intended for retinal image analysis.	(Med-IE) Image segmentation, disease grading.	Explainability
MIMIC IV	Curated Real-World Clinical Dataset	Over 300,000 hospital admissions from Beth Israel Deaconess Medical Center covering de-identified EHR data including demographics, vital signs, medications, diagnoses, and clinical notes	(Med-IE / Med-QA / Med-Gen) Used for tasks such as medical code prediction, patient outcome forecasting, clinical summarization, and question answering	Explainability

Table 1: This table provides a structured comparison of datasets used in studies on trust in LLMs for healthcare. The datasets are categorized by data type (e.g., web-scraped, curated domain-specific, synthetic, real-world, or private datasets), content (e.g., medical literature, patient records, clinical guidelines, QA pairs), task (e.g., clinical decision support, medical question-answering, document summarization, biomedical fact-checking, chatbot training), and dimensions of trustworthiness (e.g., truthfulness, privacy, safety, robustness, fairness, bias, explainability). This comparison highlights how each dataset contributes to the development of trustworthy LLMs in medical AI.

C Comparison of Models

We systematically gathered and analyzed 81 models relevant to studies on trust in LLMs for healthcare. Table 2 provides a comprehensive summary of the LLMs evaluated in these studies, detailing key aspects such as model name, release year, openness, architecture, and the institution responsible for its development. Additionally, it specifies the primary task each model is designed for, including medical question-answering, clinical decision support, and biomedical text summarization. To further assess their reliability, we categorize the models based on the dimensions of trustworthiness they address, such as truthfulness, privacy, safety, robustness, fairness and bias, and explainability. This structured overview offers valuable insights into how different LLMs are designed and evaluated to enhance trust in healthcare AI applications.

Models	Release Year	Institution	Openness	Architecture	Primary Task	Dimensions
SciBERT	2019	Allen Institute for AI	Open-source	Encoder-only	Pre-trained language model specialized for scientific text, particularly biomedical and computer science literature.	Fairness and Bias
PaLM-2	2023	Google	Closed-source	Decoder-only	Multilingual language understanding and generation, with a focus on reasoning and coding tasks.	Fairness and Bias
Mixtral-8x70B	2023	Mistral AI	Open-source	Decoder-only	Ensemble of language models aimed at improving performance across diverse language tasks.	Fairness and Bias, Safety
Med-PaLM	2023	Google Health	Closed-source	Decoder-only	Specializing in healthcare-related question answering, clinical diagnosis support, and medical literature interpretation.	Fairness and Bias
Med-PaLM 2	2024	Google Health	Closed-source	Encoder-decoder	Updated version of Med-PaLM, further improving healthcare-related tasks with enhanced accuracy and reliability in medical information retrieval, clinical reasoning, and decision support.	Fairness and Bias
Llama-13B	2023	Meta	Open-source	Decoder-only	Designed for natural language understanding and generation tasks, such as text summarization, machine translation, and conversational AI.	Fairness and Bias
XLNet	2019	Google Research	Open-source	Encoder-only	It is used for text classification, question answering, and language modeling tasks.	Fairness and Bias
DeBERTa	2020	Microsoft Research	Open-source	Encoder-only	Improves BERT and RoBERTa by enhancing the attention mechanism. It performs well in a variety of NLP tasks, such as sentence classification, question answering, and named entity recognition.	Fairness and Bias
Llama-7B	2023	Meta	Open-source	Decoder-only	Focused on general-purpose natural language understanding and generation, with potential fine-tuning for specific domains like medicine, law, and technology.	Fairness and Bias, Truthfulness
Llama 2 70Bchat	2023	Meta Platforms	Open-source	Decoder-only	Open-source conversational AI model designed for dialogue and instruction-following tasks.	Fairness and Bias, Truthfulness, Safety, Robustness,
GPT-3.5	2022	OpenAI	Closed-source	Decoder-only	Enhanced language processing capabilities, building upon GPT-3.	Fairness and Bias, Truthfulness, Safety, Robustness, Privacy
GPT2	2019	OpenAI	Open-source	Decoder-only	Text generation	Fairness and Bias, Robustness

Models	Release Year	Institution	Openness	Architecture	Primary Task	Dimensions
PMC Llama 13B	2023	Allen Institute for AI	Open-source	Decoder-only	Specialized in medical literature understanding and generation.	Fairness and Bias, Robustness
GPT-4	2023	OpenAI	Closed-source	Decoder-only	Advanced language generation and understanding across various domains.	Fairness and Bias, Safety, Robustness, Explainability, Privacy
BERT	2018	Google AI Language	Open-source	Encoder-only	Pre-trained Transformer model for a wide range of NLP tasks, such as text classification, NER, QA, etc.	Fairness and Bias, Safety, Robustness, Truthfulness
LLAMA 2 CHAT	2023	Meta AI	Open-source	Decoder-only	Language modeling	Robustness, Explainability
MEDALPACA (7B)	2023	medalpaca	Open-source	Decoder-only	Medical domain language model fine-tuned for question-answering and medical dialogue tasks.	Robustness, Privacy
CLINICAL CAMEL (13B)	2023	the AI and healthcare community	Open-source	Decoder-only	Fine-tuned for clinical applications. It is designed to assist with tasks like medical text classification, clinical decision support, information extraction from medical records, and answering clinical questions.	Robustness
GPT-2 XL	2019	OpenAI	Open-source	Decoder-only	Large-scale language model for text generation and understanding.	Robustness
T5-Large	2020	Google Research	Open-source	Encoder-decoder	It treats all NLP tasks as text-to-text tasks, meaning both the input and output are in the form of text, and it's used for tasks like translation, summarization, and question answering.	Robustness
claude-3.5-sonnet	2024	Anthropic	Closed-source	Decoder-only	It is a variant of Claude, specialized in tasks such as conversational AI, creative writing, poetry generation, and other text-based applications.	Robustness
OpenBioLLM-70B	2024	OpenBioAI	Open-source	Decoder-only	It is designed to handle tasks such as biological information extraction, gene sequence analysis, protein folding predictions, and other bioinformatics applications.	Robustness
BioMistral-7B	2023	Mistral AI	Open-source	Decoder-only	Focused on biomedical and healthcare-related text. Its tasks include medical question answering, clinical document analysis, and medical text summarization.	Robustness
Medllama3-v20	2024	MedAI Labs	Open-source	Decoder-only	Designed to assist in healthcare tasks like clinical reasoning, medical question answering, and patient record analysis.	Robustness

Models	Release Year	Institution	Openness	Architecture	Primary Task	Dimensions
ASCLEPIUS (7B)	2023	Asclepius AI	Open-source	Decoder-only	Developed for clinical and medical applications, specializing in tasks like diagnosing medical conditions from symptoms, medical text summarization, and extracting structured information from clinical documents.	Robustness, Explainability
ALPACA (7B)	2023	Stanford University	Open-source	Decoder-only	Fine-tuned version of the LLaMA model aimed at providing high-quality responses to questions, with an emphasis on maintaining ethical and accurate conversational capabilities in diverse domains.	Robustness
Google's Bard	2023	Google	Closed-source	Encoder-decoder	Conversational AI tool, focused on providing detailed, accurate, and creative responses to user queries. It can handle a variety of tasks, including web search, content generation, and complex QA.	Robustness
Text- Davinci-003	2022	OpenAI	Closed-source	Decoder-only	It is an advanced variant of GPT-3. It is designed for a wide range of natural language understanding and generation tasks, such as answering questions, summarizing text, creative writing, translation, and code generation.	Robustness, Truthfulness
LLaMa 2-7B	2023	Meta (formerly Facebook AI Research)	Open-source	Decoder-only	Designed to be a general-purpose AI for a wide range of tasks such as text generation, question answering, and summarization, with specific fine-tuning for medical and technical domains.	Robustness, Truthfulness, Privacy
ChatGPT	2022	OpenAI	Closed-source	Decoder-only	Conversational AI	Robustness, Truthfulness, Explainability, Privacy
Llama-3.1	2024	Meta AI	Open-source	Decoder-only	Multilingual large language model designed for a variety of natural language processing tasks.	Safety, privacy
ClinicalCamel-70b	2023	the AI and healthcare community	Open-source	Decoder-only	Medical language model designed for clinical research applications.	Safety, Explainability
Med42-70b	2023	M42 Health	Open-source	Decoder-only	Clinical large language model providing high-quality answers to medical questions.	Safety, Explainability
GPT-4o	2024	OpenAI	Closed-source	Decoder-only	Multimodal large language model capable of processing and generating text, audio, and images in real time.	Safety, Privacy, Explainability
Mistral	2023	Mistral AI	Open-source	Decoder-only	Language model optimized for code generation and reasoning tasks.	Safety, Robustness, Explainability
Meditron (7) (70b)	2023	École Polytechnique Fédérale de Lausanne (EPFL)	Open-source	Decoder-only	Medical language model fine-tuned for clinical decision support and medical reasoning.	Safety, Robustness, Explainability

Models	Release Year	Institution	Openness	Architecture	Primary Task	Dimensions
Claude-2.1	2023	Anthropic	Closed-source	Decoder-only	General-purpose language model for a wide range of natural language understanding and generation tasks.	Safety, Robustness
GPT-J	2021	EleutherAI	Open-source	Decoder-only	Open-source language model for text generation and understanding.	Safety, Robustness
Vicuna	2023	UC Berkeley and Microsoft Research	Open-source	Decoder-only	Conversational AI	Safety, Robustness, Truthfulness
Medalpaca-13b	2023	medalpaca	Open-source	Decoder-only	Medical domain language model fine-tuned for question-answering and medical dialogue tasks.	Safety, Truthfulness, Privacy
GPT-3	2020	OpenAI	Closed-source	Decoder-only	Natural language understanding and generation	Truthfulness, Explainability
ALBERT	2019	Google Research	Open-source	Encoder-only	Lighter version of BERT that reduces parameters for efficiency while maintaining performance. It excels in tasks such as text classification, named entity recognition, and question answering.	Truthfulness
RoBERTa	2019	Facebook AI Research	Open-source	Encoder-only	Optimized variant of BERT that removes the Next Sentence Prediction task and trains with more data and for longer periods. It is used for tasks like question answering, sentiment analysis, and text classification.	Truthfulness
BlueBERT	2019	NIH and Stanford University	Open-source	Encoder-only	BERT-based model pre-trained on clinical and biomedical text. It is designed for healthcare-related tasks, including clinical text classification, named entity recognition, and medical question answering.	Truthfulness
ClinicalBERT	2019	University of Pennsylvania	Open-source	Encoder-only	Variant of BERT fine-tuned on clinical texts, tailored for clinical NLP tasks like named entity recognition, clinical event extraction, and question answering in the medical domain.	Truthfulness
TAPAS	2020	Google Research	Open-source	Encoder-only	Designed for answering questions based on tabular data. It is used for tasks like extracting structured information from tables and processing queries in tabular datasets.	Truthfulness
LLaMA-2 13B	2023	Meta	Open-source	Decoder-only	Advanced variant of Meta's LLaMA series, designed for text generation, question answering, summarization, and other NLP tasks.	Truthfulness, Explainability, Privacy
MPT	2023	MosaicML	Open-source	Decoder-only	General-purpose LLM for text generation, summarization, language understanding, and reasoning tasks. Fine-tuned for downstream applications such as chatbot development, code generation, and other NLP tasks.	Truthfulness

Models	Release Year	Institution	Openness	Architecture	Primary Task	Dimensions
BLIP2	2023	Salesforce	Open-source	Encoder-decoder	Bootstrapping language-image pre-training, designed to bridge vision-language models with large language models for improved visual understanding and generation.	Truthfulness
InstructBLIP-7b/13b	2023	Salesforce	Open-source	Encoder-decoder	Visual instruction-tuned versions of BLIP-2, utilizing Vicuna-7B and Vicuna-13B language models, respectively, to enhance vision-language understanding through instruction tuning.	Truthfulness
LLaVA1.5-7b/13b	2023	Microsoft	Open-source	Encoder-decoder	Large language and vision assistant models with 7B and 13B parameters, respectively, designed for multimodal tasks by integrating visual information into language models.	Truthfulness
mPLUGOwl2	2023	Zhejiang University	Open-source	Encoder-decoder	Multimodal pre-trained language model designed to handle various vision-language tasks, including image captioning and visual question answering.	Truthfulness
XrayGPT	2023	University of Toronto	Open-source	Decoder-only	Specialized model for generating radiology reports from chest X-ray images, aiming to assist in medical image interpretation.	Truthfulness
MiniGPT4	2023	King Abdullah University of Science and Technology	Open-source	Decoder-only	A lightweight multimodal model designed to align vision and language models efficiently, facilitating tasks like image captioning and visual question answering.	Truthfulness
RadFM	2023	Stanford University	Open-source	Decoder-only	Foundation model tailored for radiology, focusing on interpreting medical images and integrating findings with clinical language models.	Truthfulness
Alpaca-LoRA	2023	Stanford University	Open-source	Decoder-only	It focuses on achieving good performance in tasks such as question answering and personalized dialogue.	Truthfulness
Robin- medical	2023	Robin Health	Open-source	Decoder-only	Fine-tuned for medical applications, including clinical decision support, medical question answering, and health record analysis.	Truthfulness
Flan-T5	2021	Google Research	Open-source	Encoder-decoder	Optimized for tasks like question answering, text summarization, and sentence classification, across a variety of domains.	Truthfulness, Explainability
BioBERT	2019	Korea University	Open-source	Encoder-only	Biomedical language representation learning, enhancing performance on tasks like named entity recognition, relation extraction, and question answering within the biomedical domain.	Truthfulness
Falcon Instruct (7B and 40B)	2023	Technology Innovation Institute (TII), UAE.	Open-source	Decoder-only	Instruction-tuned language model designed to follow user instructions effectively.	Truthfulness, Robustness

Models	Release Year	Institution	Openness	Architecture	Primary Task	Dimensions
Mistral Instruct (7B)	2023	Mistral AI	Open-source	Decoder-only	Instruction-tuned language model designed to follow user instructions effectively.	Truthfulness, Robustness
Falcon	2023	Technology Innovation Institute (TII), UAE.	Open-source	Decoder-only	General-purpose language model optimized for text understanding, generation, question answering, and reasoning tasks. Focused on efficient deployment for industry-scale applications.	Truthfulness, Robustness
LLaVA-Med	2024	Microsoft	Open-source	Encoder-decoder	Large language and vision assistant for biomedicine, trained to handle visual instruction tasks in the biomedical field, aiming for capabilities similar to GPT-4.	Truthfulness, Explainability
Claude-3	2024	Anthropic	Closed-source	Decoder-only	General-purpose LLM (QA, dialogue, reasoning, summarization)	Explainability
GPT-4o-mini	2024	OpenAI	Closed-source	Decoder-only	Natural language processing (NLP), text generation, and understanding.	Explainability
ASCLEPIUS (13B)	2023	Asclepius AI	Open-source	Decoder-only	Medical NLP, clinical text analysis, and healthcare-related tasks.	Explainability
MedViLaM	2023	Cite	Open-source	Encoder-decoder	Medical vision-language tasks, combining image and text analysis for healthcare.	Explainability
Med-MoE	2023	Cite	Open-source	Decoder-only	Medical NLP, leveraging Mixture of Experts (MoE) for specialized healthcare tasks.	Explainability
Gemini Pro	2023	Google DeepMind	Closed-source	Decoder-only	Multi-modal NLP, combining text, image, and other data types for advanced AI tasks	Explainability
Gemini-1.5	2024	Google DeepMind	Closed-source	Decoder-only	Multimodal reasoning, long-context understanding, QA, generation	Explainability
AlpaCare (7B) (13B)	2023	Cite	Open-source	Decoder-only	Healthcare-focused NLP, clinical text analysis, and medical decision support	Explainability
Yi (6B)	2023	01.AI (China)	Open-source	Decoder-only	General-purpose NLP, text generation, and fine-tuning for specific applications.	Explainability

Models	Release Year	Institution	Openness	Architecture	Primary Task	Dimensions
Phi-2 (2.7B)	2023	Microsoft	Open-source	Decoder-only	Lightweight NLP, text generation, and fine-tuning for specific tasks.	Explainability
SOLAR (10.7B)	2023	Upstage AI	Open-source	Decoder-only	General-purpose NLP, text generation, and fine-tuning for specific domains.	Explainability
InternLM2 (7B)	2023	Shanghai AI Laboratory (China)	Open-source	Decoder-only	General-purpose NLP, text generation, and fine-tuning for specific applications.	Explainability
Llama3-(8B and 70B)	2024	Meta	Open-source	Decoder-only	General-purpose NLP, text generation, and fine-tuning for specific applications.	Privacy, Explainability
CodeLlama-(7B, 13B, and 34B)	2023	Meta	Open-source	Decoder-only	Code generation, code completion, and programming assistance.	Privacy
Mixtral-8x7B and 8x22B	2023	Mistral AI	Open-source	Decoder-only	General-purpose NLP, text generation, and fine-tuning for specific domains.	Privacy
Qwen-(7B, 14B, 32B, 72B)-Chat	2023	Alibaba	Open-source	Decoder-only	Chat-oriented NLP, conversational AI, and text generation.	Privacy
GLM-4	2024	Tsinghua University	Open-source	Encoder-decoder	Advanced NLP, text generation, and multi-modal tasks.	Privacy

Table 2: Detailed Comparison of GPT Models Evaluated for Trust in Healthcare LLMs, Including Model Name, Release Year, Institution, Openness, Architecture, Primary Tasks (e.g., Medical Question-Answering, Clinical Decision Support, Biomedical Text Summarization, Medical Report Generation), and Key Trustworthiness Dimensions (Truthfulness, Privacy, Safety, Robustness, Fairness and Bias, Explainability).