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 has the potential to revolutionize clinical decision-making, medical research, and patient care. As LLMs are increasingly integrated into healthcare systems, several critical challenges must be addressed to ensure their reliable and ethical deployment. These challenges include truthfulness, where models generate misleading information; privacy, with risks of unintentional data retention; robustness, requiring defenses against adversarial attacks; fairness, addressing biases in clinical outcomes; explainability, ensuring transparent decision-making; and safety, mitigating risks of misinformation and medical errors. Recently, researchers have begun developing benchmarks and evaluation frameworks to systematically assess the trustworthiness of LLMs. However, the **trustworthiness of LLMs in healthcare** remains underexplored, lacking a systematic review that provides a comprehensive understanding and future insights into this area. This **survey** bridges this gap by providing a comprehensive overview of the recent research of existing methodologies and solutions aimed at mitigating the above risks in healthcare. By focusing on key trustworthiness dimensions including truthfulness, privacy and safety, robustness, fairness and bias, and explainability, we present a thorough analysis of how these issues impact the reliability and ethical use of LLMs in healthcare. This paper highlights ongoing efforts and offers insights into future research directions to ensure the safe and trustworthy deployment of LLMs in healthcare.

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, thereby safeguarding patient safety and supporting clinical decision-making. Similarly, *fairness and bias* must be addressed to prevent discriminatory patterns in model predictions, 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. The lack of transparency in model reasoning complicates clinical adoption and raises concerns about accountability.

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 datasets, solutions, and methodologies aimed at improving the trustworthiness of LLMs in healthcare.

2 Datasets, Models, and Tasks

We first conducted an extensive search for papers on the trustworthiness of LLMs in healthcare. Our search utilized a range of keyword combinations, including terms such as 'large language models,' 'foundation model,' 'medical,' 'clinical,' 'explainability,' 'truthfulness,' 'trustworthiness,' 'safety,' 'fairness,' 'robustness,' and 'privacy.' We explored several reputable venues, including Arxiv, PubMed, ACL, EMNLP, NAACL, ICML, NeurIPS, ICLR, KDD, Nature, Science, AAAI, and IJCAI, with a focus on recent publications post-2021. After reviewing the search results, we identified a total of 30,595 papers. Following the removal of duplicates, we narrowed the focus to 69 papers that specifically addressed the truthfulness, privacy, safety, robustness, fairness, bias, and explainability of LLMs in the healthcare domain.

We then summarized all the datasets, models, and tasks relevant to research on trust in LLMs for healthcare, providing a comprehensive overview of their applications and contributions to this domain. **The datasets** used in studies of trust in LLMs for healthcare are categorized by the dimensions of trustworthiness they address in Appendix A, 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 primary purpose 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. **The models** assessed in studies on trust in LLMs for the healthcare domain are outlined, along with their trustworthiness dimensions, in Appendix B, where we summarized key details such as the model name, release year, task, and the institution responsible for its development. **The tasks** covered various primary focuses of LLMs in healthcare. Based on insights from the survey by Liu et al. (2024), these tasks are outlined as follows:

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 state-

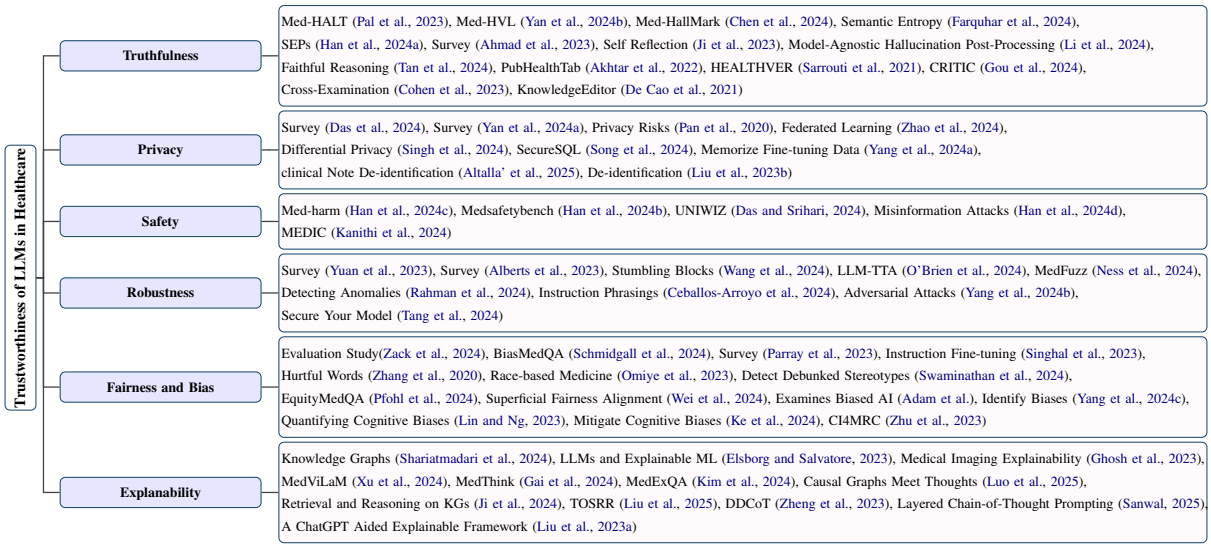

Figure 1: Summary of the recent research across various dimensions of trustworthiness of LLMs in healthcare.

ments), 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 and 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 1 provides a summary of the recent research on trust in LLMs for healthcare across key dimensions of trustworthiness.

3.1 Truthfulness

Ensuring the *truthfulness* of LLMs in healthcare is vital, as inaccurate information can significantly impact patient care and clinical outcomes. Given their influence on diagnoses and treatment decisions, it is essential to develop effective methods to detect and mitigate hallucinations and factual inaccuracies.

Hallucinations in medical LLMs arise from reliance on unverified sources, biases in training data, and limitations in contextual understanding and sequential reasoning (Ahmad et al., 2023). Addressing these issues requires robust evaluation frameworks, self-correction mechanisms, and uncertainty quantification techniques.

The Med-HALT benchmark (Pal et al., 2023) is designed to evaluate hallucinations in medical LLMs by using reasoning-based tests like 'False Confidence' and 'None of the Above,' as well as memory-based tests to assess how well the model recalls medical knowledge. On the other hand, the interactive self-reflection methodology (Ji et al., 2023) aims to reduce hallucinations in medical question-answering tasks by introducing an iterative feedback loop where the model refines its responses through self-evaluation and knowledge adjustment.

In the context of multimodal models, Med-HVL (Yan et al., 2024b) introduces two key metrics—Object Hallucination and Domain Knowledge Hallucination—to quantify hallucinations in Large Vision-Language Models (LVLMs). This framework also uses the CHAIR (Caption Hallucination Assessment with Image Relevance) metric to assess object hallucinations in image captioning. Med-HallMark (Chen et al., 2024) takes it a step further by providing a multi-task evaluation framework with a hierarchical categorization of hallucinations, introducing the MediHall Score for assessing hallucination severity and the MediHallDetector, a multitask-trained LVLM for hallucination detection.

Researchers have also investigated semantic entropy, a probabilistic measure of uncertainty, to

detect hallucinations in LLMs. For example, Farquhar et al. (2024) leverage semantic entropy to identify confabulations—hallucinations where the model generates arbitrary or incorrect outputs. While effective, this approach is computationally expensive, limiting its scalability. To overcome this, Han et al. (2024a) introduce Semantic Entropy Probes (SEPs), which approximate semantic entropy directly from hidden states. By eliminating the need for multiple output samples, SEPs significantly reduce computational overhead. Both methods have been successfully applied to biomedical datasets, such as BioASQ.

Although these techniques offer valuable contributions, hallucination mitigation methods often lack adaptability, being either task-specific or requiring expensive retraining. To address this gap, MEDAL (Li et al., 2024) introduces a model-agnostic post-processing framework that integrates with any medical summarization model. MEDAL uses a self-examining correction model to improve factual accuracy without adding extra computational costs, providing a practical solution to the issue of medical hallucinations.

Collectively, these studies underscore the multifaceted challenge of ensuring truthfulness in medical LLMs. By leveraging benchmarking frameworks, self-correction mechanisms, and entropy-based uncertainty measures, researchers can develop complementary strategies for detecting, mitigating, and quantifying hallucinations. A key focus of these efforts is the development of quantifiable scoring methods, enabling systematic assessment and comparison across different models. These evaluation techniques not only help identify the most reliable and effective LLMs for healthcare applications but also provide actionable insights for further improvements.

Factual accuracy is fundamental to building trust in LLMs, especially in healthcare, where reliable and verifiable information is critical. However, current LLMs lack effective mechanisms to trace claims back to their original sources, underscoring the urgent need for improved validation techniques to ensure safe and trustworthy medical applications. To address these challenges, several studies have introduced innovative approaches to enhance the transparency, accuracy, and reliability of healthcare LLMs.

Tan et al. (2024) propose an approach that integrates multiple perspectives from scientific literature to evaluate conflicting arguments, thereby improving LLM reasoning. Similarly, Akhtar et al. (2022) introduce PubHealthTab, a table-based dataset designed for validating public health claims against noisy evidence, while Sarrouti et al. (2021) present HEALTHVER, a dataset tailored for evidence-based fact-checking of health-related claims. These structured benchmarks provide a foundation for assessing and refining the reliability of LLM-generated medical information.

Beyond dataset-driven validation, self-correction mechanisms have been explored to improve LLM truthfulness. Gou et al. (2024) introduce CRITIC, a framework inspired by human fact-checking practices, enabling LLMs to validate and refine their responses through iterative feedback and evaluation. Expanding on automated fact-checking, Cohen et al. (2023) propose a cross-examination framework, where an examiner LLM identifies inconsistencies through multi-turn interactions with the original model. Unlike fully automated verification pipelines, CRITIC incorporates human-like evaluation strategies, enhancing the trustworthiness of fact-checking in medical contexts.

Overall, these studies advance the factual accuracy and transparency of LLMs in healthcare by introducing structured benchmarks, iterative validation processes, and automated fact-checking strategies. By incorporating these approaches, researchers can enhance the reliability of medical LLMs, ensuring they deliver more accurate, evidence-based insights to support clinical decision-making.

3.2 Privacy

LLM-based healthcare applications pose significant *privacy* risks due to their ability to memorize and reproduce sensitive patient data (Das et al., 2024). Unauthorized data exposure can lead to confidentiality breaches, ethical concerns, and compliance violations (Pan et al., 2020). Addressing these risks requires privacy safeguards at different stages of model development.

A major challenge is unintended data retention and leakage, where LLMs memorize fine-tuning data, increasing re-identification risks. Studies show that domain-specific LLMs, such as Medalpaca, can retain sensitive data, making privacy breaches more likely (Yang et al., 2024a). Additionally, adversarial attacks like prompt injection and inference attacks can further exploit these vulnerabilities, as demonstrated by the SecureSQL benchmark (Song et al., 2024).

To mitigate these risks, pre-training privacy safeguards focus on de-identification. Altalla' et al. (2025) assess GPT-3.5 and GPT-4 in clinical note de-identification and synthetic data generation. Similarly, Liu et al. (2023b) propose a GPT-4-enabled framework for masking private information while maintaining text structure. However, de-identification remains imperfect, as attackers may infer sensitive details from anonymized text.

During fine-tuning, techniques such as federated learning (Zhao et al., 2024) and differential privacy (Singh et al., 2024), as highlighted by Liu et al. (2024), play a crucial role in safeguarding patient data. Federated learning enables decentralized training without sharing raw data, but it demands high computational resources. Differential privacy adds noise to protect sensitive information but can reduce model accuracy.

Adversarial defenses remain limited. The SecureSQL benchmark (Song et al., 2024) highlights LLM vulnerabilities to structured query attacks. While chain-of-thought (COT) prompting offers partial mitigation, it does not eliminate the risk of data exposure.

Researchers address privacy concerns in healthcare LLMs through two primary approaches: de-identification, which alters real data to prevent re-identification, and synthetic data generation, which creates artificial data to eliminate reliance on sensitive patient information. While these strategies enhance privacy protection and maintain model effectiveness, challenges remain in long-term memorization control and adversarial robustness, requiring further research to strengthen data security and prevent unintended information retention.

3.3 Safety

Ensuring the *safety* of LLMs in healthcare is critical, as these models must not generate harmful responses. A key safety concern, noted by Han et al. (2024d), is that modifying just 1.1 % of a model's weights can embed persistent biomedical inaccuracies while maintaining overall performance. This highlights the need for rigorous validation mechanisms and safety assessments to prevent misleading medical information before clinical use.

To systematically assess safety, MedSafetyBench (Han et al., 2024b) was introduced as the first benchmark designed to evaluate LLM safety in medical contexts. It includes 1,800 harmful medical queries alongside safety-optimized responses generated using advanced LLMs and adversarial techniques. Results indicate that publicly available medical LLMs fail to meet safety standards, but fine-tuning with MedSafetyBench significantly improves safety without compromising performance.

A major challenge is that adversarial actors can manipulate LLMs to generate unsafe outputs, while excessive safety alignment may induce hallucinations. To address this, Das and Srihari (2024) propose UNIWIZ, a two-step framework that unifies safety alignment and factual knowledge retrieval. Their safety-priming approach synthesizes safety-focused training data, while a retrieval mechanism ensures that model outputs remain factually accurate. Models fine-tuned on UNIWIZ outperform larger state-of-the-art instruction-tuned models across multiple safety and accuracy metrics.

Another key contribution to safety alignment is from Han et al. (2024c), who provide the first comprehensive safety evaluation for medical LLMs (MedLLMs). They define key concepts of medical safety and alignment and introduce Med-Harm, a dataset designed to evaluate both general and medical-specific risks. This dataset assesses how well LLMs handle harmful medical questions, ensuring they adhere to safety and ethical standards in medical AI.

Further advancing safety assessments, Kanithi et al. (2024) introduce MEDIC, a multi-dimensional trustworthiness evaluation framework that systematically assesses medical LLMs across critical dimensions of clinical competence including medical reasoning and clinical safety.

These studies collectively offer a multi-faceted approach to LLM safety in healthcare. MedSafetyBench (Han et al., 2024b) provides a standardized benchmark for safety evaluation and fine-tuning, while UNIWIZ (Das and Srihari, 2024) introduces a structured framework that prevents hallucinations while reinforcing safety. Han et al. (2024c) focus on comprehensive safety alignment, establishing Med-Harm to evaluate domain-specific risks. Lastly, MEDIC (Kanithi et al., 2024) offers a holistic evaluation franmework for improving practical application of LLMs in clinical settings. Together, these efforts contribute to a more rigorous and systematic framework for assessing and improving LLM safety in medical applications.

3.4 Robustness

Enhancing the *robustness* of LLMs is crucial for their reliability in healthcare applications. A key approach involves developing adversarial test samples

tailored to the medical domain, such as synthetic anomaly cases (Yuan et al., 2023) and boundary stress testing (Wang et al., 2024), to assess model resilience. However, creating clinically meaningful adversarial samples presents unique challenges, as Alberts et al. (2023) highlight the need to align adversarial testing methods with the complexities of real-world medical data, where medical dependencies must be accounted for.

In addition to adversarial testing, uncertainty quantification is another important avenue to improve robustness. LLM-TTA (O'Brien et al., 2024) explores test-time adaptation techniques to enhance model performance on rare or unfamiliar cases, which are common in medical diagnostics. Unlike adversarial robustness, which focuses on resistance to manipulated inputs, uncertainty quantification aims to identify when models are likely to be incorrect, providing a complementary safety mechanism.

Another critical question is whether benchmark performance truly reflects medical robustness. MedFuzz (Ness et al., 2024) challenges assumptions in MedQA by modifying questions to test if models rely on rigid, dataset-specific patterns rather than genuine clinical reasoning. This research exposes vulnerabilities in LLMs, revealing that subtle changes in input can significantly impact performance, raising concerns about their reliability in dynamic medical environments.

Instruction robustness is also a growing concern. Ceballos-Arroyo et al. (2024) examine how variations in medical instructions affect performance across different LLMs, finding that specialized medical models may be more fragile than general-purpose models when instructions are reworded. This counterintuitive result suggests that excessive domain adaptation may decrease flexibility and reduce robustness.

Adversarial vulnerabilities also pose direct security risks. Yang et al. (2024b) investigate two adversarial attack strategies across medical tasks using real patient data, demonstrating that fine-tuned models are especially vulnerable to poisoning attacks that subtly alter learned weights. While adversarial data does not always degrade general performance, it can introduce dangerous biases into specific medical predictions, making early attack detection a priority.

To protect against adversarial manipulations, Tang et al. (2024) introduce Secure Your Model, a framework that strengthens LLM robustness with cryptographic prompt authentication. This mechanism ensures that only verified and secure prompts are processed, mitigating vulnerabilities associated with prompt injections and adversarial attacks, and reducing the risk of model exploitation in medical contexts.

While all these studies address LLM robustness, they differ in their primary focus. MedFuzz (Ness et al., 2024) and Ceballos-Arroyo et al. (2024) expose vulnerabilities in existing benchmarks and task instructions, questioning whether current evaluation methods truly measure robustness or merely reflect dataset biases. In contrast, Yang et al. (2024b) and Alberts et al. (2023) highlight adversarial threats, demonstrating how medical LLMs can be subtly manipulated, raising concerns about their security in real-world applications. Meanwhile, LLM-TTA (O'Brien et al., 2024) takes a different approach, focusing on uncertainty quantification rather than adversarial resistance to enhance reliability in handling unfamiliar cases. Secure Your Model (Tang et al., 2024) provides an additional security layer by introducing proactive adversarial defenses through prompt protection mechanisms, ensuring resilience against manipulation risks.

These studies highlight that robustness is a multifaceted challenge, requiring advancements in evaluation methods and defensive mechanisms. Ensuring LLMs can handle adversarial scenarios, integrate domain knowledge, and adapt to language variations is crucial for their safe deployment in healthcare. Strengthening robustness through testing and resilience-building enhances the trustworthiness of medical LLMs, making them more reliable in complex clinical settings.

3.5 Fairness and Bias

Ensuring *fairness* in LLMs is crucial in healthcare, where biased models can result in unequal treatment outcomes. Research has highlighted biases in clinical data and practice related to race, gender, and disability. For example, Omiye et al. (2023) examine the potential for harmful or inaccurate race-based content in LLMs, while Zack et al. (2024) discusse how language models encode societal biases that can affect healthcare outcomes. These studies underscore the need for fairness in LLM development to ensure equitable healthcare delivery.

Efforts to address bias focus on both detection and mitigation. Swaminathan et al. (2024) offers an automated method for detecting race-based medicine stereotypes, while BiasMedQA

(Schmidgall et al., 2024) benchmarks cognitive biases in medical tasks across multiple models, revealing varying bias resilience. Mitigation strategies, such as bias education, and one-shot and few-shot bias demonstrations, are proposed to reduce but not fully eliminate bias. Pfohl et al. (2024) introduce frameworks for assessing health equity-related harms in LLMs, including EquityMedQA, a dataset for equity-focused testing. Additionally, Wei et al. (2024) distinguishe between intrinsic fairness, rooted in model training, and behavioral fairness, which relates to model operation in real-world applications, advocating for both to ensure equitable outcomes.

Bias is also explored in closed-source models. Zack et al. (2024) evaluate racial and gender biases in clinical scenarios, finding that models often amplify societal biases. Similarly, Adam et al. show that biased AI recommendations can affect emergency decisions, while Yang et al. (2024c) identify healthcare disparities in model predictions based on patient demographics. For open-source LLMs, techniques such as reinforcement learning with clinician feedback (Zack et al., 2024) and data augmentation during pre-training (Parray et al., 2023) enhance training data quality to reduce bias. In contrast, for closed-source models, where internal representations are inaccessible, strategies like instruction fine-tuning (Singhal et al., 2023) and prompt engineering (Schmidgall et al., 2024) are employed to improve fairness in outputs.

Further, Zhang et al. (2020) investigate how LLM embeddings can encode biases, particularly in clinical tasks, and applies adversarial debiasing to mitigate disparities. Lin and Ng (2023) identify cognitive biases in BERT, while Ke et al. (2024) use a multi-agent framework to explore how LLMs can mitigate biases in clinical decision-making. Additionally, Zhu et al. (2023) introduce CI4MRC, a method that addresses name-related bias in Machine Reading Comprehension (MRC) tasks by applying a causal interventional paradigm.

In addressing bias, solutions vary depending on whether the bias is at the individual level (e.g., name-related information) or dataset level (e.g., biased racial distributions). Open-sourced LLMs benefit from direct accessibility, enabling robust interventions to mitigate bias. In contrast, closed-sourced models, due to their inaccessibility, rely on methods such as instruction fine-tuning and external post-processing tools to refine model outputs and reduce bias.

3.6 Explanability

The lack of *explainability* in LLMs poses a significant barrier to building trust with clinical practitioners, thereby restricting their adoption in real-world healthcare systems. To address this challenge, efforts in clinical practice and research have prioritized improving the transparency of LLMs by developing more effective explanation mechanisms and incorporating human oversight. For instance, Shariatmadari et al. (2024) enhance biomedical applications by integrating knowledge graphs with language models and visualizing attention probabilities to provide clear and interpretable explanations for model predictions. Elsborg and Salvatore (2023) employ local explanation models to generate intuitive, case-specific insights, further advancing the explainability of LLMs and fostering trust in their medical applications.

Another critical area of research focuses on medical imaging explainability. Ghosh et al. (2023) propose a method that iteratively decomposes a black-box (BB) model into interpretable expert models and a residual network, where expert models specialize in specific data subsets and explain their reasoning using First-Order Logic (FOL).

Efforts to improve reasoning and multimodal integration in LLMs have also gained momentum. MedViLaM (Xu et al., 2024) and MedThink (Gai et al., 2024) focus on improving medical understanding and reasoning by integrating both textual and visual data for complex medical tasks such as question answering and medical image classification, aiming to provide a more holistic view for decision-making. In contrast, MedExQA (Kim et al., 2024) prioritizes explainability in medical QA systems by providing multiple explanations for its responses. Causal Graphs Meet Thoughts (Luo et al., 2025) and Retrieval and Reasoning on KGs (Ji et al., 2024) explore how to enhance complex reasoning in LLMs through knowledge graphs (KGs), helping these models retrieve relevant information from structured sources to answer questions more effectively and with better justification.

For more specialized applications, TOSRR (Liu et al., 2025) introduce tree-organized self-reflective retrieval, which aims to improve performance in the niche area of Traditional Chinese Medicine (TCM), combining LLMs with self-reflection techniques for more accurate and contextually relevant responses. Studies like DDCoT (Zheng et al., 2023) and Layered Chain-of-Thought Prompting (San-

wal, 2025) develop methods for multimodal reasoning across multiple agents, with DDCoT focusing on chain-of-thought prompting for structured reasoning across tasks. Lastly, frameworks like A ChatGPT Aided Explainable Framework (Liu et al., 2023a) aim for zero-shot diagnosis and interpretation, ensuring more accessible AI support in real-world medical settings.

The trend towards explainability enhancement encompasses both intrinsic and post-hoc approaches. Intrinsic methods, such as knowledge graphs and decomposed expert models, integrate explainability directly into the models, offering natural transparency. Post-hoc methods, including local explanations and textual justifications, elucidate model decisions after predictions, improving user comprehension and trust.

4 Future Directions

We have reviewed key trust challenges in LLMs and existing solutions. This section highlights current limitations and proposes future directions.

While efforts to enhance LLM *truthfulness* in healthcare have advanced, gaps remain, including limited adaptability in hallucination mitigation and weak source attribution. Future research should prioritize model-agnostic post-processing, improved self-correction, enhanced uncertainty quantification, and real-time fact-checking via structured knowledge integration.

Existing *privacy* safeguards, such as de-identification and federated learning, remain imperfect. Strengthening de-identification methods, reinforcing federated learning defenses, refining differential privacy, and exploring homomorphic encryption and real-time audits are crucial next steps.

Safety remains a pressing concern due to adversarial attacks and excessive safety alignment-induced hallucinations. Future work should focus on robust validation mechanisms, refined safety alignment strategies, and comprehensive evaluation frameworks.

Improving *robustness* remains critical, with adversarial testing and uncertainty quantification methods needing to better handle medical data complexities. Future research should focus on clinically relevant adversarial tests, enhancing uncertainty techniques, and improving instruction robustness.

Despite progress in *fairness*, key gaps persist, including the need for comprehensive bias assessments and real-world testing of mitigation strategies. Standardizing fairness metrics, conducting real-world evaluations, and assessing long-term impacts on healthcare equity are critical for progress.

Advancements in *explainability* have yet to bridge significant gaps. Future research should focus on integrating explainability into clinical workflows, developing interactive explanations, and improving multimodal integration to enhance transparency and trust.

Recently, multi-agent frameworks like TriageAgent (Lu et al., 2024) have been introduced to streamline complex clinical tasks through agent collaboration. By harnessing their capabilities, we can embed trustworthiness mitigation and evaluation within a multi-agent system, enabling proactive monitoring and intervention across key areas such as truthfulness, privacy, robustness, fairness, and explainability in healthcare LLMs.

Addressing these gaps will ensure that LLMs can be effectively integrated into healthcare systems, improving their reliability, privacy, safety, fairness, and transparency.

5 Conclusion

The integration of LLMs into healthcare holds great promise, but realizing their full potential requires addressing the critical challenges outlined in this survey. A key concern is truthfulness, as inaccuracies in medical LLMs pose serious risks to patient safety, making their detection and mitigation an ongoing research priority. Equally vital are privacy, safety, robustness, fairness, and explainability, ensuring responsible deployment in real-world clinical settings.

While existing solutions show progress, much work remains to enhance the reliability, transparency, and ethical implications of LLMs in healthcare. Future research must refine these areas, balancing performance with trustworthiness while preventing LLM-based systems systems from worsening healthcare disparities. Comprehensive benchmarks, cross-disciplinary collaboration, and model accountability frameworks will be essential. Additionally, regulatory oversight and ethical guidelines must ensure LLM applications align with medical standards and patient rights.

Ultimately, achieving safe and equitable AI-driven healthcare will require ongoing efforts to improve both technical capabilities and societal frameworks.

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*.

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.

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

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*.

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.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonas Elsborg and Marco Salvatore. 2023. Using llms and explainable ml to analyze biomarkers at single-cell level for improved understanding of diseases. *Biomolecules*, 13(10):1516.

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. 2024. Medthink: Explaining medical visual question answering via multimodal decision-making rationale. *arXiv preprint arXiv:2404.12372*.

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*.

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, 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.

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.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3843–3860.

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.

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.

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.

Chang Liu, Ying Chang, Jianmin Li, Yiqian Qu, Yu Li, Lingyong Cao, and Shuyuan Lin. 2025. Improving tcm question answering through tree-organized self-reflective retrieval with llms. *Preprint*, arXiv:2502.09156.

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. 2024. A survey on medical large language models: Technology, application, trustworthiness, and future directions. *arXiv preprint arXiv:2406.03712*.

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.

Hang Luo, Jian Zhang, and Chujun Li. 2025. Causal graphs meet thoughts: Enhancing complex reasoning in graph-augmented llms. *arXiv preprint arXiv:2501.14892*.

Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E Priebe, and Eric Horvitz. 2024. Medfuzz: Exploring the robustness of large language models in medical question answering. *arXiv preprint arXiv:2406.06573*.

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.

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*.

Manish Sanwal. 2025. Layered chain-of-thought prompting for multi-agent llm systems: A comprehensive approach to explainable large language models. *arXiv preprint arXiv:2501.18645*.

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.

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.

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*.

Lijian Xu, Hao Sun, Ziyu Ni, Hongsheng Li, and Shaoting Zhang. 2024. Medvilam: A multimodal large language model with advanced generalizability and explainability for medical data understanding and generation. *arXiv preprint arXiv:2409.19684*.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024a. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*.

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

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, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024b. Adversarial attacks on large language models in medicine. *arXiv preprint arXiv:2406.12259*.

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024c. 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.

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.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

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

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.

A Comparison of Datasets

We systematically collected and analyzed 41 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	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.	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	Replicate common clinically relevant cognitive biases	Fairness and Bias
NEJM Healer	Real Data and Curated Domain-Specific Dataset,	Consists of Clinical Cases, Diagnostic Pathways, Educational Materials, Interactive Learning Modules	Tasks including Diagnostic Skill Development, Medical Education, Simulated Decision-Making, Feedback and Improvement	Fairness and Bias
EquityMedQA	Curated domain-specific datasets and synthetic data	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.	Evaluate the performance of LLMs in generating unbiased, equitable medical responses.	Fairness and Bias
"bias" medical dataset	curated, domain-specific dataset comprising real-world healthcare data from a large U.S. health system	Includes patient demographics, medical histories, and healthcare utilization records.	Evaluate and identify racial biases in algorithms used for healthcare management.	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.	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	Real Data and Curated Domain-Specific Datasets	Consists of electronic health records include patient Demographics, Clinical Data, Medical Notes, Treatment Records, Time-series Data	Various medical and machine-learning tasks, including Clinical Decision Support, Disease Modeling, Natural Language Processing, Time-series Analysis, Education and Research	Fairness and Bias, Explainability
MedQA	Curated Domain-Specific Datasets	A benchmark that includes questions drawn from the United States Medical License Exam (USMLE).	Exam the physicians to test their ability to make clinical decisions	Fairness and Bias, Robustness, Explainability
PMC-Patients	Curated dataset derived from public text corpora.	Contains 167,000 patient summaries extracted from 141,000 PMC articles	Designed to benchmark ReCDS systems through two primary tasks: Patient-to-Article Retrieval (PAR), Patient-to-Patient Retrieval (PPR)	Robustness
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.	Epidemiological studies, clinical decision-rule improvement, machine learning in healthcare.	Robustness

Datasets	Data Type	Content	Task	Dimensions
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).	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
UNIWIZ	Synthetic and curated data, including: 17,638 quality-controlled 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.	Fine-tune large language models to enhance their performance in generating safe and knowledge-grounded conversations.	Safety
SciFact	Curated Domain-Specific Dataset.	Includes claims and corresponding evidence abstracts, each annotated with labels indicating whether the claim is supported or refuted, along with rationales justifying the decision.	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.	Facilitates evidence-based fact-checking by providing claims and corresponding evidence tables for verification.	Truthfulness
LAMA	Curated Domain-Specific Dataset.	Comprises a set of knowledge sources, each containing a collection of facts.	To probe pretrained 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.	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	consists of real anonymized queries from Google's search engine users, paired with answers derived from entire Wikipedia articles.	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.	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.	Fact extraction and verification, where models are trained to determine the veracity of claims based on provided evidence.	Truthfulness
HEALTHVER	Curated Domain-Specific Dataset.	Contains 14,330 evidence-claim pairs (SUPPORTS, REFUTES, NEUTRAL) on health claims, mainly COVID-19, verified against scientific articles.	Used to train and evaluate models in verifying health claims by classifying them based on scientific evidence.	Truthfulness

Datasets	Data Type	Content	Task	Dimensions
Med-HALT	Synthetic and Real Data, Curated Domain-Specific Dataset, and Public Dataset	Consist of Reasoning-Based Assessments, Memory-Based Assessments, Medical Scenarios, Evaluation Metrics	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)	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.	Task including Medical Image Captioning, Text-Image Retrieval, Medical Reasoning	Truthfulness
Med-HallMark	Curated Domain-Specific Dataset, , Synthetic and Real Data (includes a mix of real-world medical data and synthetically generated hallucination scenarios)	Diverse medical multimodal data, including text, images, and paired annotations. Hierarchically categorized hallucination data, addressing both structural (e.g., object-level) and contextual (e.g., domain knowledge) hallucinations.	Task including Hallucination Detection, Hallucination Evaluation, Mitigation Analysis	Truthfulness
BioASQ	Curated Domain-Specific Dataset; Real Data.	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.	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.	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.	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.	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	Medical question answering, focusing on consumer health questions received by the U.S. National Library of Medicine.	Truthfulness
MASH-QA	Curated Domain-Specific Dataset.	Approximately 25,000 question-answer pairs sourced from WebMD, covering a wide range of healthcare topics.	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.	Evaluate and analyze data leakage risks in LLMs, particularly concerning SQL query generation and execution.	Privacy

Datasets	Data Type	Content	Task	Dimensions
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	Designed to enhance large language models (LLMs) for medical applications	Privacy
Electronic Health Records (EHR) at (KHCC)	Private dataset	gpt-3.5-turbo	Clinical research, outcome analysis.	Privacy
MedVQA	Curated domain-specific dataset	A collection of medical visual question answering pairs, designed to train and evaluate models that interpret medical images and answer related questions.	Visual question answering, medical image understanding.	Explainability
MedExQA	Curated domain-specific dataset	A dataset focused on medical examination questions and answers, intended to aid in the development of AI models for medical exam preparation and assessment.	Question answering, educational assessment.	Explainability
MedMCQA	Curated domain-specific dataset	A multiple-choice question-answering dataset in the medical domain, aimed at training models to handle medical examinations and practice questions.	Multiple-choice question answering, medical education.	Explainability
TCM Medical Licensing Examination(MLE)	Curated domain-specific dataset	A dataset comprising questions and answers from Traditional Chinese Medicine licensing examinations.	Educational assessment, question answering.	Explainability
Pneumonia Dataset	Curated domain-specific dataset	Medical images (such as chest X-rays) labeled for the presence or absence of pneumonia, used for training diagnostic models.	Image classification, disease detection.	Explainability
Montgomery Dataset	Curated domain-specific dataset	Chest X-ray images with manual segmentations of the lung fields, useful for pulmonary research.	Image segmentation, tuberculosis detection.	Explainability
Shenzhen Dataset	Curated domain-specific dataset	Chest X-ray images collected in Shenzhen, China, with annotations for tuberculosis manifestations.	Disease classification, image analysis.	Explainability
IDRID Dataset	Curated domain-specific dataset	Retinal images with annotations for diabetic retinopathy lesions, intended for retinal image analysis.	Image segmentation, disease grading.	Explainability
BrainTumor Dataset	Curated domain-specific dataset	MRI images of brain tumors with corresponding labels, used for developing diagnostic and segmentation models.	Tumor detection, image segmentation.	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.

B Comparison of Models

We systematically gathered and analyzed 78 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, 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	Primary Task	Dimensions
SciBERT	2019	Allen Institute for AI	Pre-trained language model specialized for scientific text, particularly biomedical and computer science literature.	Fairness and Bias
PaLM-2	2023	Google	Multilingual language understanding and generation, with a focus on reasoning and coding tasks.	Fairness and Bias
Mixtral-8x70B	2023	Mistral AI	Ensemble of language models aimed at improving performance across diverse language tasks.	Fairness and Bias, Safety
Med-PaLM	2023	Google Health	Specializing in healthcare-related question answering, clinical diagnosis support, and medical literature interpretation.	Fairness and Bias
Med-PaLM 2	2024	Google Health	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	Designed for natural language understanding and generation tasks, such as text summarization, machine translation, and conversational AI.	Fairness and Bias
XLNet	2019	Google Research	It is used for text classification, question answering, and language modeling tasks.	Fairness and Bias
DeBERTa	2020	Microsoft Research	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	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 conversational AI model designed for dialogue and instruction-following tasks.	Fairness and Bias, Truthfulness, Safety, Robustness,
GPT-3.5	2022	OpenAI	Enhanced language processing capabilities, building upon GPT-3.	Fairness and Bias, Truthfulness, Safety, Robustness, Privacy
GPT2	2019	OpenAI	Text generation	Fairness and Bias, Robustness
PMC Llama 13B	2023	Allen Institute for AI	Specialized in medical literature understanding and generation.	Fairness and Bias, Robustness

Models	Release Year	Institution	Primary Task	Dimensions
GPT-4	2023	OpenAI	Advanced language generation and understanding across various domains.	Fairness and Bias, Safety, Robustness, Explainability, Privacy
BERT	2018	Google AI Language	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	Language modeling	Robustness, Explainability
MEDALPACA (7B)	2023	medalpaca	Medical domain language model fine-tuned for question-answering and medical dialogue tasks.	Robustness, Privacy
CLINICAL CAMEL (13B)	2023	the AI and healthcare community	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	Large-scale language model for text generation and understanding.	Robustness
T5-Large	2020	Google Research	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	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	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	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	Designed to assist in healthcare tasks like clinical reasoning, medical question answering, and patient record analysis.	Robustness
ASCLEPIUS (7B)	2023	Asclepius AI	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	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

Models	Release Year	Institution	Primary Task	Dimensions
Google's Bard	2023	Google	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	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)	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	Conversational AI	Robustness, Truthfulness, Explainability, Privacy
Llama-3.1	2024	Meta AI	Multilingual large language model designed for a variety of natural language processing tasks.	Safety
ClinicalCamel-70b	2023	the AI and healthcare community	Medical language model designed for clinical research applications.	Safety, Explainability
Med42-70b	2023	M42 Health	Clinical large language model providing high-quality answers to medical questions.	Safety, Explainability
GPT-4o	2024	OpenAI	Multimodal large language model capable of processing and generating text, audio, and images in real time.	Safety, Privacy, Explainability
Mistral	2023	Mistral AI	Language model optimized for code generation and reasoning tasks.	Safety, Robustness, Explainability
Meditron (7) (70b)	2023	École Polytechnique Fédérale de Lausanne (EPFL)	Medical language model fine-tuned for clinical decision support and medical reasoning.	Safety, Robustness, Explainability
Claude-2.1	2023	Anthropic	General-purpose language model for a wide range of natural language understanding and generation tasks.	Safety, Robustness
GPT-J	2021	EleutherAI	Open-source language model for text generation and understanding.	Safety, Robustness
Vicuna	2023	UC Berkeley and Microsoft Research	Conversational AI	Safety, Robustness, Truthfulness

Models	Release Year	Institution	Primary Task	Dimensions
Medalpaca-13b	2023	medalpaca	Medical domain language model fine-tuned for question-answering and medical dialogue tasks.	Safety, Truthfulness, Privacy
GPT-3	2020	OpenAI	Natural language understanding and generation	Truthfulness, Explainability
ALBERT	2019	Google Research	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	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	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	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	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	Advanced variant of Meta's LLaMA series, designed for text generation, question answering, summarization, and other NLP tasks.	Truthfulness, Explainability, Privacy
MPT	2023	MosaicML	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
BLIP2	2023	Salesforce	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	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	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	Multimodal pre-trained language model designed to handle various vision-language tasks, including image captioning and visual question answering.	Truthfulness

Models	Release Year	Institution	Primary Task	Dimensions
XrayGPT	2023	University of Toronto	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	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	Foundation model tailored for radiology, focusing on interpreting medical images and integrating findings with clinical language models.	Truthfulness
Alpaca-LoRA	2023	Stanford University	It focuses on achieving good performance in tasks such as question answering and personalized dialogue.	Truthfulness
Robin- medical	2023	Robin Health	Fine-tuned for medical applications, including clinical decision support, medical question answering, and health record analysis.	Truthfulness
Flan-T5	2021	Google Research	Optimized for tasks like question answering, text summarization, and sentence classification, across a variety of domains.	Truthfulness
BioBERT	2019	Korea University	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.	Instruction-tuned language model designed to follow user instructions effectively.	Truthfulness, Robustness
Mistral Instruct (7B)	2023	Mistral AI	Instruction-tuned language model designed to follow user instructions effectively.	Truthfulness, Robustness
Falcon	2023	Technology Innovation Institute (TII), UAE.	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	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
GPT-4o-mini	2024	OpenAI	Natural language processing (NLP), text generation, and understanding.	Explainability
ASCLEPIUS (13B)	2023	Asclepius AI	Medical NLP, clinical text analysis, and healthcare-related tasks.	Explainability

Models	Release Year	Institution	Primary Task	Dimensions
MedViLaM	2023	Xu et al. (2024)	Medical vision-language tasks, combining image and text analysis for healthcare.	Explainability
Med-MoE	2023	Jiang et al. (2024)	Medical NLP, leveraging Mixture of Experts (MoE) for specialized healthcare tasks.	Explainability
Gemini Pro	2023	Google DeepMind	Multi-modal NLP, combining text, image, and other data types for advanced AI tasks	Explainability
AlpaCare (7B) (13B)	2023	Zhang et al. (2023)	Healthcare-focused NLP, clinical text analysis, and medical decision support	Explainability
Yi (6B)	2023	01.AI (China)	General-purpose NLP, text generation, and fine-tuning for specific applications.	Explainability
Phi-2 (2.7B)	2023	Microsoft	Lightweight NLP, text generation, and fine-tuning for specific tasks.	Explainability
SOLAR (10.7B)	2023	Upstage AI	General-purpose NLP, text generation, and fine-tuning for specific domains.	Explainability
InternLM2 (7B)	2023	Shanghai AI Laboratory (China)	General-purpose NLP, text generation, and fine-tuning for specific applications.	Explainability
Llama3-(8B and 70B)	2024	Meta	General-purpose NLP, text generation, and fine-tuning for specific applications.	Privacy
CodeLlama-(7B, 13B, and 34B)	2023	Meta	Code generation, code completion, and programming assistance.	Privacy
Mixtral-8x7B and 8x22B	2023	Mistral AI	General-purpose NLP, text generation, and fine-tuning for specific domains.	Privacy
Qwen-(7B, 14B, 32B, 72B)-Chat	2023	Alibaba	Chat-oriented NLP, conversational AI, and text generation.	Privacy
GLM-4	2024	Tsinghua University	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, 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).