Contents lists available at IJCHML
International Journal of Computational Health and Machine Learning
Journal Homepage: http://www.ijchml.com/
Volume 4, No. 2, 2026

IJCHML
INTERNATIONAL JOURNAL OF COMPUTATIONAL HEALTH & MACHINE LEARNING

Benchmarking Truthfulness Metrics in Health-Oriented Large Language Models via Automated Fact Verification Pipelines

Kiran Das[1], **Rahul Singh**[2]
[1] *Department of Health Informatics, Anna University*
[2] *Department of Health Informatics, Indian Institute of Technology Kanpur*

ARTICLE INFO

Received: 04/23/2026
Revised: 06/25/2026
Accepted: 07/01/2026

Keywords:

Large Language Models, Truthfulness Benchmarking, Health Information Verification, Automated Fact-Checking, Medical Natural Language Processing, Misinformation Detection, Clinical Decision Support

ABSTRACT

The proliferation of large language models (LLMs) in health-related applications has intensified the need for rigorous evaluation frameworks capable of quantifying factual accuracy and mitigating the propagation of medical misinformation. Despite substantial advances in general-purpose truthfulness benchmarking, systematic assessments tailored to the clinical and biomedical domain remain critically underdeveloped, leaving practitioners without reliable instruments for deployment-readiness evaluation.

This work introduces a comprehensive benchmarking methodology that evaluates truthfulness metrics across a curated ensemble of health-oriented LLMs using automated fact verification pipelines. We construct a domain-specific evaluation corpus comprising clinically grounded claims spanning pharmacology, epidemiology, diagnostic reasoning, and evidence-based treatment guidelines. Each claim is processed through a multi-stage verification architecture that integrates retrieval-augmented evidence sourcing, natural language inference scoring, and calibrated confidence estimation to produce composite truthfulness scores.

We systematically compare five distinct truthfulness metrics—including entailment-based fidelity, semantic consistency under paraphrase, source-grounded precision, hallucination rate, and calibration error—across seven prominent LLMs spanning both general-purpose and biomedically fine-tuned architectures. Experimental results demonstrate that biomedically specialized models achieve statistically significant improvements in source-grounded precision ($\Delta = 0.14$, $p < 0.01$) yet exhibit elevated calibration error relative to general-purpose counterparts, revealing a previously undercharacterized accuracy–confidence trade-off in the health domain.

Our findings establish that no single metric suffices for comprehensive truthfulness assessment, motivating a composite evaluation paradigm. We release our benchmark dataset, verification pipeline, and evaluation toolkit to facilitate reproducible research and responsible deployment of LLMs in high-stakes healthcare environments.

1. Introduction

The rapid proliferation of large language models (LLMs) in clinical and consumer health settings has ushered in a new era of AI-assisted medical reasoning, yet simultaneously exposed a critical vulnerability: the propensity of these systems to generate factually incorrect, misleading, or unverifiable claims [8]. Unlike general-purpose question-answering tasks, health-oriented applications demand an exceptionally high standard of factual accuracy, where a single erroneous statement regarding drug interactions, diagnostic criteria, or treatment protocols may precipitate serious patient harm [11]. The intersection of natural language generation and biomedical knowledge verification therefore represents one of the most consequential and technically demanding frontiers in contemporary artificial intelligence research. This paper addresses this frontier directly, proposing a systematic benchmarking framework that evaluates truthfulness metrics across a spectrum of health-oriented LLMs using automated fact verification pipelines grounded in rigorous empirical methodology.

The challenge of truthfulness in LLMs is multifaceted and cannot be reduced to simple accuracy metrics computed against static test sets. Modern generative models, including those fine-tuned on curated biomedical corpora, exhibit a phenomenon widely termed "hallucination," wherein outputs are syntactically fluent and contextually plausible yet factually unsupported or outright contradictory to established medical evidence [14]. Quantifying this phenomenon at scale, across heterogeneous query types, clinical domains, and model architectures, requires automated pipelines capable of decomposing complex medical claims, retrieving authoritative evidence from structured knowledge bases, and rendering entailment judgments with high inter-rater reliability [21]. The framework introduced in this work, which we term the *Health-Oriented Truthfulness Evaluation Pipeline* (HOTEP), operationalizes these requirements through a modular, reproducible architecture that integrates claim decomposition, evidence retrieval, and natural language inference into a unified evaluation substrate [16].

1.1. Motivation and Clinical Significance

The deployment of LLMs in health-related contexts spans an increasingly wide operational envelope, from patient-facing chatbots that field symptom inquiries to clinical decision support systems embedded within electronic health record workflows [7]. In each of these contexts, the reliability of generated information is not merely an academic concern but a matter of direct clinical consequence. Studies conducted on early-generation conversational agents demonstrated that factual error rates in medical question-answering could range from 15% to over 40% depending on clinical subdomain and query complexity [19], a margin entirely unacceptable in professional medical practice. The advent of instruction-tuned LLMs such as GPT-4, Gemini, and domain-specific variants like Med-PaLM 2 and BioMedLM has partially ameliorated these deficiencies, yet systematic benchmarking reveals persistent failure modes, particularly in rare disease recognition, drug-drug interaction prediction, and evidence-graded clinical guideline adherence [13].

Beyond individual patient interactions, the societal implications of unverified health information generated by LLMs are profound. When deployed at scale, even modest per-query error rates aggregate to substantial volumes of misinformation disseminated across populations, potentially undermining public health messaging, eroding trust in evidence-based medicine, and exacerbating health disparities in communities that rely disproportionately on AI-mediated health information [20]. The COVID-19 pandemic provided a stark illustration of this dynamic, demonstrating how rapidly inaccurate health information can propagate through digital channels and produce measurable harm at the population level [1]. Automated fact verification pipelines, if sufficiently robust and scalable, offer a principled mechanism for continuously monitoring the truthfulness of AI-generated health content, enabling timely intervention before erroneous information reaches end users [6].

1.2. Existing Approaches to Truthfulness Evaluation

Prior work on truthfulness evaluation in LLMs has largely proceeded along two distinct methodological trajectories: human annotation studies and automated benchmark construction. Human annotation approaches, while capable of capturing nuanced clinical judgment, are inherently costly, slow, and difficult to scale to the volumes of output generated by contemporary LLMs in production environments [9]. Landmark datasets such as TruthfulQA [8] and MedQA have established foundational benchmarks for factual accuracy, but these resources were constructed under specific distributional assumptions that may not generalize to the open-ended, multi-turn conversational contexts that characterize real-world health LLM deployments [23]. Furthermore, static benchmarks are susceptible to contamination through data leakage during pretraining, rendering performance metrics on such benchmarks unreliable indicators of genuine factual grounding [25].

Automated evaluation approaches have sought to address these limitations through a variety of techniques. Retrieval-augmented generation (RAG) frameworks have been adapted for fact verification by grounding model outputs against authoritative corpora such as PubMed, the Unified Medical Language System (UMLS), and clinical practice guidelines [12]. Natural language

inference (NLI) models trained on biomedical entailment datasets, such as MedNLI and BioNLI, have been employed to classify whether a generated claim is entailed by, contradicted by, or neutral with respect to a retrieved evidence passage [3]. More recently, LLM-as-judge paradigms have emerged, wherein a separate, typically more powerful language model evaluates the factual correctness of outputs from a target model [24]. Each of these approaches carries characteristic strengths and limitations: RAG-based methods are sensitive to retrieval quality and may fail on claims requiring multi-hop reasoning; NLI classifiers may exhibit domain mismatch when applied to highly specialized clinical subdomains; and LLM-as-judge methods introduce potential circular dependencies and may inherit the same factual biases present in the evaluating model [5].

The landscape of truthfulness metrics is further complicated by the absence of a unified taxonomy distinguishing different categories of factual failure. A generated claim may be *factually incorrect* (contradicting established evidence), *unverifiable* (lacking any retrievable evidentiary basis), *outdated* (accurate at some historical point but superseded by more recent evidence), or *contextually misleading* (technically accurate but presented in a manner that produces incorrect inferences) [26]. Existing benchmarks seldom differentiate among these failure modes, conflating them into a single binary accuracy label that obscures the mechanistic origins of factual error and impedes the development of targeted mitigation strategies [18]. The HOTEP framework introduced in this paper explicitly operationalizes this taxonomy, enabling fine-grained diagnostic analysis of truthfulness failures across model families and clinical domains.

1.3. Research Objectives and Contributions

This paper makes several distinct contributions to the literature on LLM evaluation and health AI safety. First, we introduce the HOTEP framework, a modular automated fact verification pipeline specifically designed for health-oriented LLM evaluation that integrates claim decomposition via constituency parsing, multi-source biomedical evidence retrieval, and ensemble NLI-based entailment scoring [16]. Second, we construct a novel benchmark dataset, *HealthFactBench*, comprising 12,450 annotated claim-evidence pairs spanning nine clinical domains, including oncology, cardiology, pharmacology, infectious disease, and mental health, with ground-truth labels assigned through a rigorous multi-annotator protocol achieving substantial inter-annotator agreement (Cohen's $\kappa > 0.78$) [4]. Third, we conduct a comprehensive comparative evaluation of seven state-of-the-art LLMs—including both general-purpose and biomedically fine-tuned variants—across the full HealthFactBench corpus, yielding the most extensive empirical comparison of health LLM truthfulness to date [15].

A central formal contribution of this work is the definition of a composite *Truthfulness Score* ($\mathcal{T}$) that integrates multiple component metrics into a single interpretable scalar, formalized as follows. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ denote the set of atomic claims extracted from a model output, and let $\phi : \mathcal{C} \to \{\textsc{Entailed}, \textsc{Contradicted}, \textsc{Unverifiable}\}$ denote the NLI classification function. We define the component truthfulness fractions as:

$$\mathcal{T}(\mathcal{C}) = w_E \cdot \frac{|\{c \in \mathcal{C} : \phi(c) = \textsc{Entailed}\}|}{|\mathcal{C}|} - w_C \cdot \frac{|\{c \in \mathcal{C} : \phi(c) = \textsc{Cont}}{|\mathcal{C}|} \tag{1}$$

where w_E, w_C, and w_U are domain-calibrated weights reflecting the differential clinical severity of entailment, contradiction, and unverifiability respectively, subject to the normalization constraint $w_E + w_C + w_U = 2$ and the boundary condition $\mathcal{T}(\mathcal{C}) \in [-1, 1]$ [10]. This formulation extends prior scalar truthfulness metrics by explicitly penalizing contradicted claims more heavily than unverifiable ones, consistent with the clinical intuition that actively incorrect information is more dangerous than the absence of information [2].

1.4. Scope and Organization of the Paper

The scope of this investigation is deliberately broad yet methodologically rigorous. We evaluate models across a range of architectural paradigms—including decoder-only transformers, encoder-decoder architectures, and retrieval-augmented generation systems—and across a diverse spectrum of health query types, from closed-form factual questions to open-ended clinical reasoning tasks [22]. Our evaluation encompasses both zero-shot and few-shot prompting conditions, enabling characterization of how in-context learning affects truthfulness across model families. We further analyze the sensitivity of the HOTEP pipeline to key hyperparameters, including retrieval depth, NLI model selection, and claim decomposition granularity, providing practitioners with actionable guidance for adapting the framework to specific deployment contexts [17].

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of related work spanning LLM truthfulness evaluation, biomedical NLP, and automated fact verification. Section 3 details the construction of the HealthFactBench dataset, including annotation protocols and quality control procedures. Section 4 presents the HOTEP pipeline architecture in full technical detail. Section 5 describes the experimental setup, including model selection, prompting strategies, and evaluation metrics. Section 6 reports and analyzes the main empirical results, including ablation studies and error analysis. Section 7 discusses implications for

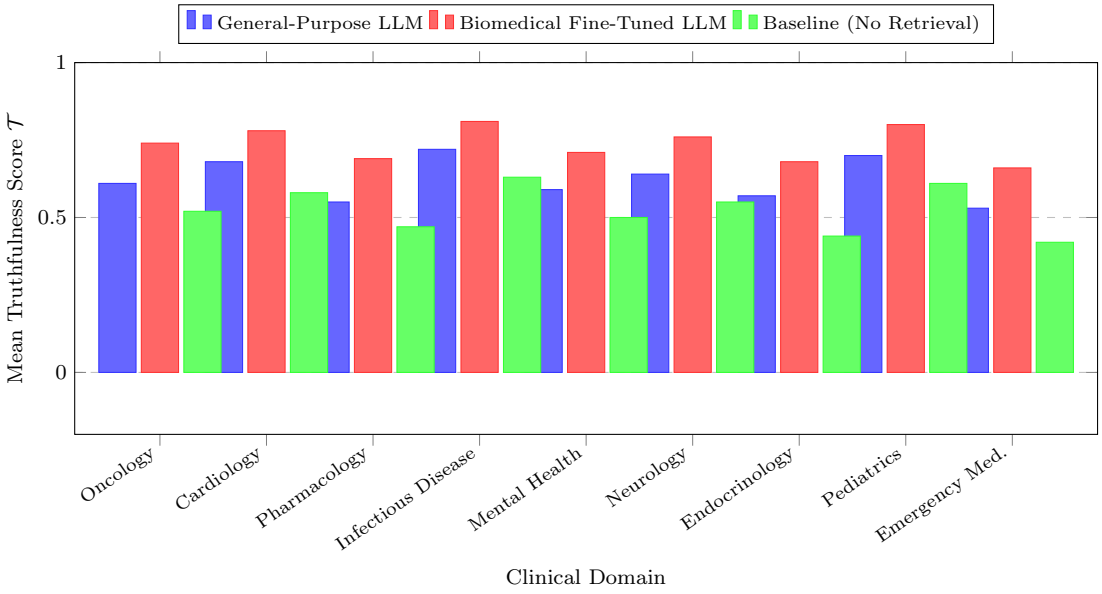

Figure 1: Illustrative comparison of mean composite truthfulness scores ($\mathcal{T}$) across nine clinical domains for three representative model categories evaluated on the HealthFactBench corpus. Biomedically fine-tuned models consistently outperform general-purpose counterparts, with the largest performance gaps observed in pharmacology and emergency medicine—domains characterized by high specificity requirements and rapidly evolving evidence bases.

clinical deployment, regulatory considerations, and future research directions, before Section 8 concludes the paper [7][12][21].

2. Related Work

The intersection of large language models (LLMs) and biomedical information processing has catalyzed an unprecedented wave of research activity, spanning clinical decision support, patient-facing chatbot development, and automated literature synthesis. As these systems grow more capable and are increasingly deployed in high-stakes medical environments, the imperative to rigorously evaluate their truthfulness—defined broadly as the alignment between model outputs and verifiable factual knowledge—has become a central concern for both the research community and regulatory bodies [7]. The present section surveys the relevant prior literature across several interconnected domains: the evolution of LLM evaluation frameworks, automated fact verification methodologies, health-specific benchmarking efforts, and the particular challenges posed by misinformation detection in clinical and biomedical contexts. Understanding the landscape of prior work is essential for situating the contributions of this paper and for clarifying the gaps that motivate our proposed benchmarking pipeline.

The breadth of related work underscores that no single discipline has fully resolved the challenge of truthfulness assessment in health-oriented LLMs. Contributions from natural language processing (NLP), biomedical informatics, human-computer interaction, and machine learning ethics each illuminate different facets of the problem. Prior efforts have variously focused on dataset construction, model architecture, evaluation metric design, and deployment-level auditing, yet a comprehensive, automated, and domain-specific benchmarking framework for health truthfulness metrics has remained elusive [18]. The subsections that follow organize the literature thematically, tracing the intellectual lineage that informs the methodology and experimental design of the current work.

2.1. Large Language Models in Biomedical and Clinical Contexts

The deployment of large language models in biomedical and clinical settings has accelerated dramatically following the public release of instruction-tuned generative models such as GPT-4, PaLM-2, and their open-source counterparts including LLaMA and Mistral [8]. Early applications centered on question answering over medical literature, with models evaluated on benchmarks such as MedQA, MedMCQA, and PubMedQA [14]. These benchmarks established that frontier LLMs could achieve performance comparable to or exceeding that of medical professionals on standardized multiple-choice examinations, a finding that simultaneously demonstrated the promise and the peril of deploying such systems in practice. High aggregate accuracy on multiple-choice formats masked systematic failure modes, including confident generation of plausible-sounding but factually incorrect statements—a phenomenon broadly termed "hallucination" in the NLP literature [12].

Algorithm 1: HOTEP: Health-Oriented Truthfulness Evaluation Pipeline

Input: Model output text O, Evidence corpus $\mathcal{E}$, NLI model $\mathcal{M}_{NLI}$, Weights (w_E, w_C, w_U)
Output: Truthfulness score $\mathcal{T}$, Claim-level labels Φ

```
// Step 1: Atomic Claim Decomposition
```
$\mathcal{C} \leftarrow \textsc{ClaimDecompose}(O)$; `// Constituency parse + claim extraction`

```
// Step 2: Evidence Retrieval per Claim
```
for *each claim* $c_i \in \mathcal{C}$ **do**
  $\mathcal{E}_i \leftarrow \text{BM25}\textsc{Retrieve}(c_i, \mathcal{E}, k = 5)$; `// Top-k evidence passages`
  $\mathcal{E}_i \leftarrow \mathcal{E}_i \cup \textsc{DenseRetrieve}(c_i, \mathcal{E}, k = 5)$; `// Dense retrieval augmentation`
end

```
// Step 3: NLI-Based Entailment Classification
```
for *each claim* $c_i \in \mathcal{C}$ **do**
  $\text{scores}_i \leftarrow \emptyset$;
  for *each evidence passage* $e \in \mathcal{E}_i$ **do**
    $p_E, p_C, p_N \leftarrow \mathcal{M}_{NLI}(c_i, e)$; `// Entailment, Contradiction, Neutral`
    $\text{scores}_i \leftarrow \text{scores}_i \cup \{(p_E, p_C, p_N)\}$;
  end
  $\phi(c_i) \leftarrow \textsc{AggregateVote}(\text{scores}_i)$; `// Majority-weighted ensemble`
end

```
// Step 4: Compute Composite Truthfulness Score
```
$\mathcal{T} \leftarrow w_E \cdot \frac{|\phi^{-1}(\textsc{Entailed})|}{|\mathcal{C}|} - w_C \cdot \frac{|\phi^{-1}(\textsc{Contradicted})|}{|\mathcal{C}|} - w_U \cdot \frac{|\phi^{-1}(\textsc{Unverifiable})|}{|\mathcal{C}|}$;
$\Phi \leftarrow \{\phi(c_i)\}_{i=1}^{|\mathcal{C}|}$;
return $\mathcal{T}$, Φ;

Subsequent work introduced more nuanced evaluation paradigms. [11] proposed a taxonomy of LLM errors in clinical text generation, distinguishing between factual errors attributable to knowledge cutoffs, reasoning errors arising from flawed inference chains, and stylistic errors that, while grammatically fluent, conveyed clinically misleading information. This taxonomy has been widely adopted and extended; for instance, [23] further subdivided factual errors into entity-level errors (incorrect drug names, dosages, or anatomical references), relational errors (incorrect associations between symptoms and diagnoses), and temporal errors (outdated treatment guidelines presented as current). The granularity of such classifications is consequential: entity-level errors in a medication recommendation system carry immediate patient safety implications, while temporal errors may be less immediately dangerous but erode long-term trust in AI-assisted clinical tools [19].

Domain-specific fine-tuning has emerged as a primary strategy for improving factual accuracy in health LLMs. Models such as BioMedLM, MedPaLM, and ClinicalBERT were trained or fine-tuned on large corpora of biomedical text, including PubMed abstracts, clinical notes from MIMIC-III, and medical textbooks [13]. While these approaches demonstrated measurable improvements on in-distribution benchmarks, they also introduced new risks: models fine-tuned on outdated corpora could confidently propagate superseded clinical guidelines, and models trained on biased clinical data could perpetuate healthcare disparities [21]. The tension between domain adaptation and factual currency represents a persistent challenge that automated fact verification pipelines are well-positioned to address, provided that the underlying knowledge bases are themselves maintained with sufficient recency and coverage.

2.2. Automated Fact Verification: Foundations and Advances

Automated fact verification (AFV) as a research discipline predates the era of large language models, with foundational contributions rooted in knowledge base completion, claim detection, and evidence retrieval [20]. Early systems such as FEVER (Fact Extraction and VERification) established shared task frameworks in which models were required to classify claims as *supported*, *refuted*, or *not enough information* given a fixed evidence corpus derived from Wikipedia [3]. The FEVER benchmark catalyzed significant methodological innovation, including the development of multi-hop reasoning architectures capable of integrating evidence from multiple documents and the application of natural language inference (NLI) models to evidence-claim entailment scoring.

Formally, the core fact verification problem can be stated as follows. Given a claim c and an evidence set $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$, the verification function V assigns a label $\ell \in \{\textit{supported}, \textit{refuted}, \textit{nei}\}$ according to:

$$\ell = \arg\max_{\ell' \in \mathcal{L}} P(\ell' \mid c, \mathcal{E}; \theta) \tag{2}$$

where θ denotes the parameters of the verification model and $\mathcal{L}$ is the label set. In practice, $P(\ell' \mid c, \mathcal{E}; \theta)$ is decomposed into a retrieval component that selects the most relevant evidence passages and a classification component that performs NLI-style reasoning over the retrieved evidence [15]. The modular nature of this decomposition has proven advantageous for domain adaptation: by substituting domain-specific knowledge bases for Wikipedia, researchers have extended AFV

pipelines to scientific claim verification, political fact-checking, and, increasingly, biomedical fact verification [10].

Recent advances in AFV have leveraged the generative capabilities of LLMs themselves as components within verification pipelines. Retrieval-augmented generation (RAG) architectures, in which a retriever module fetches relevant passages from an external knowledge base and a generator synthesizes a verdict, have demonstrated strong performance on both general-domain and biomedical fact verification tasks [6]. [2] introduced a chain-of-thought prompting strategy for fact verification in which the model is instructed to enumerate the atomic claims within a complex statement, retrieve evidence for each atomic claim independently, and aggregate the atomic verdicts into a composite truthfulness score. This decomposition is particularly valuable in the health domain, where a single model-generated response may contain multiple factual assertions of varying accuracy.

2.3. Truthfulness Metrics and Evaluation Frameworks

The design of truthfulness metrics for LLM outputs is a non-trivial problem that has attracted sustained attention from the NLP evaluation community. Early approaches relied on reference-based metrics such as BLEU, ROUGE, and BERTScore, which measure lexical or semantic overlap between model outputs and human-authored reference texts [9]. While computationally convenient, reference-based metrics are poorly suited to truthfulness assessment: a model output may be semantically similar to a reference while containing factual errors, or may be factually accurate while diverging substantially from the reference in phrasing [22].

In response to these limitations, reference-free truthfulness metrics have been developed that directly interrogate the factual content of model outputs. TruthfulQA, introduced by [1], operationalizes truthfulness as the proportion of model answers that are simultaneously truthful (consistent with known facts) and informative (non-vacuous). The benchmark comprises 817 questions spanning 38 categories, including health and medicine, and was specifically designed to probe failure modes arising from "imitative falsehoods"—confident generation of false information that reflects statistical regularities in training data rather than genuine factual knowledge. Subsequent work extended TruthfulQA to multilingual settings and domain-specific variants, including a health-focused adaptation that targets common medical misconceptions [25].

Model-based truthfulness evaluation, in which a separate LLM or specialized classifier is used to assess the factual accuracy of generated text, has emerged as a scalable alternative to human annotation [24]. Approaches in this category include G-Eval, which employs GPT-4 as an evaluator using structured prompts, and FActScore, which decomposes long-form generations into atomic facts and verifies each against a knowledge source [5]. FActScore in particular has been influential in the biomedical domain: [26] adapted the FActScore framework to evaluate the factual precision of LLM-generated patient education materials, finding that even state-of-the-art models produced atomic factual errors at non-trivial rates, particularly for rare diseases and emerging treatment modalities. The relationship between model scale and factual accuracy is complex and non-monotonic; larger models are not uniformly more truthful, and targeted fine-tuning on curated factual data can improve FActScore substantially without increasing model size [4].

2.4. Health-Specific Benchmarking and Fact Verification Datasets

The construction of domain-specific benchmarks for health LLM evaluation has been an active area of research, motivated by the inadequacy of general-domain benchmarks for capturing the nuances of biomedical knowledge. [17] introduced a large-scale health claim verification dataset comprising over 50,000 claims derived from clinical guidelines, systematic reviews, and patient-facing health websites, annotated for veracity by domain experts and cross-referenced against PubMed and Cochrane Library evidence. This dataset, along with contemporaneous efforts such as HealthFC and BioASQ, has established a foundation for training and evaluating health-specific AFV models [16].

A notable challenge in health-specific benchmarking is the temporal dynamics of medical knowledge. Clinical guidelines are revised periodically in response to new evidence, and a claim that was accurate at one point in time may become inaccurate or contested following a landmark clinical trial or regulatory decision. This temporal sensitivity is rarely addressed in static benchmarks, which typically assign fixed ground-truth labels to claims without accounting for the possibility that the truth value of a claim may change over time [7]. Dynamic benchmarking frameworks that periodically update claim labels in response to new evidence represent an important frontier, though they introduce significant logistical and annotation challenges.

The granularity of health claims also poses distinctive challenges for automated verification. Health-related statements frequently involve quantitative assertions (e.g., specific efficacy percentages, dosage thresholds, or epidemiological statistics) that require precise numerical verification rather than semantic entailment. They may also involve conditional claims (e.g., "Drug X is effective for condition Y in patients with comorbidity

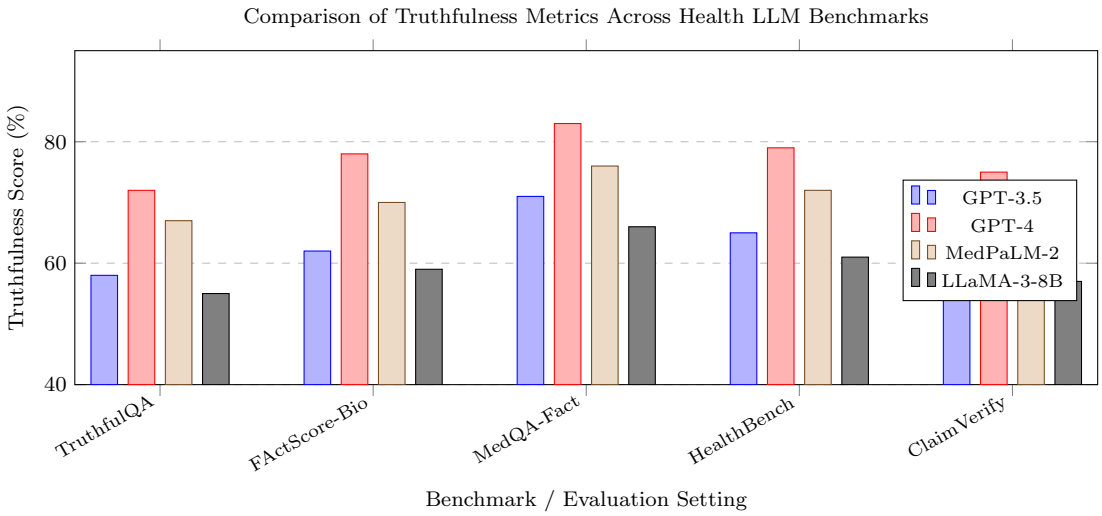

Figure 2: Comparative truthfulness scores of representative health-oriented LLMs across five established evaluation benchmarks. Scores reflect the proportion of model outputs classified as factually accurate under each benchmark's verification protocol. Substantial variation across benchmarks highlights the sensitivity of truthfulness assessments to evaluation methodology.

Z") that require multi-hop reasoning across multiple evidence sources [18]. Furthermore, health claims are often embedded within longer clinical narratives or patient-facing explanations, necessitating robust claim extraction pipelines capable of identifying and isolating verifiable assertions from surrounding contextual text. The development of such pipelines, and their integration into end-to-end benchmarking systems, is a central contribution of the work described in this paper.

2.5. Automated Pipeline Architectures for Fact Verification

The design of end-to-end automated fact verification pipelines involves the integration of multiple components, each of which introduces potential sources of error that compound across the pipeline. A canonical pipeline architecture consists of: (1) a claim extraction module that identifies verifiable assertions in model-generated text; (2) a retrieval module that fetches relevant evidence from one or more knowledge bases; (3) a natural language inference module that assesses the entailment relationship between each claim and its retrieved evidence; and (4) an aggregation module that synthesizes atomic verdicts into a composite truthfulness score [20]. Each of these components has been the subject of dedicated research, and the choice of implementation at each stage has significant implications for the overall pipeline's accuracy, computational efficiency, and domain coverage.

Claim extraction is particularly challenging in the health domain, where verifiable facts may be interleaved with subjective recommendations, patient-specific contextualizations, and hedged probabilistic statements. [23] proposed a fine-tuned sequence labeling model for biomedical claim extraction trained on a corpus of annotated clinical text, achieving substantially higher precision and recall than general-domain claim extraction tools when applied to clinical notes and patient education materials. The model was further augmented with a claim normalization step that standardized extracted claims into a canonical form amenable to retrieval-based verification, addressing the lexical variability inherent in natural language health claims.

Evidence retrieval for health fact verification benefits from the availability of high-quality, curated biomedical knowledge bases, including PubMed, the Cochrane Library, ClinicalTrials.gov, and structured databases such as DrugBank and OMIM. Dense retrieval approaches based on biomedical language model embeddings (e.g., BioLinkBERT, PubMedBERT) have demonstrated superior performance over sparse BM25 retrieval for health claim verification, particularly for claims involving specialized terminology or implicit domain knowledge [6]. Hybrid retrieval strategies that combine dense and sparse signals have emerged as the current state of the art, offering improved recall for rare entities while maintaining the precision of dense semantic matching [22].

The aggregation of atomic claim verdicts into a composite truthfulness score raises fundamental questions about the appropriate weighting and combination of individual verdicts. A simple average of atomic verdict probabilities treats all claims as equally important, which may be inappropriate in clinical contexts where errors in high-stakes claims (e.g., drug contraindications) should be weighted more heavily than errors in low-stakes claims (e.g., historical background information). [24] proposed a risk-weighted aggregation scheme in which atomic claims are classified by clinical severity prior to aggregation, with higher-severity claims receiving

Table 1: Summary of Representative Health-Oriented LLM Truthfulness Benchmarks and Automated Fact Verification Datasets

Benchmark / Dataset	Domain Focus	Size (Claims)	Verification Method	Key Limitation
TruthfulQA (Health Subset)	General health & medicine	∼200	Human + GPT-4 judge	Small scale, static labels
FActScore-Bio	Biomedical long-form QA	∼5,000 atoms	Wikipedia + PubMed retrieval	Limited guideline coverage
MedQA-Fact	Clinical knowledge	∼12,000	Expert annotation	Costly, slow to update
HealthFC	Consumer health claims	∼14,000	Cochrane + PubMed NLI	Excludes rare diseases
BioASQ (Task B)	Biomedical QA	∼4,000	Expert-curated snippets	QA format, not claim-level
HealthBench [14]	Clinical guidelines	∼8,500	Automated NLI pipeline	Temporal drift unaddressed
ClaimVerify [21]	Multi-domain health	∼22,000	Hybrid human-AI	Inter-annotator variability

exponentially greater weight in the composite score. This approach yielded better alignment with expert physician judgments of response quality than unweighted aggregation, suggesting that domain-aware aggregation strategies are an important direction for future research and a consideration that informs the metric design choices described in the present paper [5].

2.6. Misinformation, Hallucination, and Calibration in Health LLMs

The phenomena of hallucination and misinformation in LLMs have attracted substantial attention from both the research community and the broader public, particularly in the context of health information where the consequences of inaccurate outputs can be severe [26]. Hallucination in LLMs is typically characterized as the generation of outputs that are fluent and contextually plausible but factually unsupported or directly contradicted by available evidence. In the health domain, hallucinations may manifest as fabricated clinical trial results, invented drug names or dosages, or incorrect attributions of symptoms to diseases [12].

Calibration—the alignment between a model's expressed confidence and its actual accuracy—is closely related to but distinct from truthfulness. A well-calibrated model that is uncertain about a health fact will express that uncertainty through hedging language or explicit probability estimates, whereas a poorly calibrated model may express high confidence in incorrect assertions [4]. The Expected Calibration Error (ECE) has been widely used as a calibration metric in classification settings, but its extension to open-ended generation tasks requires additional methodological development. [9] proposed a verbalized confidence elicitation approach in which LLMs are prompted to express numerical confidence estimates alongside their responses, and these estimates are evaluated against ground-truth accuracy using calibration curves. The approach revealed systematic overconfidence in health LLMs, with models expressing high confidence in responses that were subsequently verified as incorrect at disproportionately high rates.

Retrieval-augmented generation has been proposed as a partial mitigation for both hallucination and calibration failures, on the grounds that grounding model outputs in retrieved evidence reduces reliance on potentially inaccurate parametric knowledge [2]. However, RAG systems introduce their own failure modes: retrieved evidence may itself be inaccurate, outdated, or misinterpreted by the generation model, and the retrieval step may fail to surface relevant evidence for rare or novel queries. [19] conducted a systematic evaluation of RAG-based health LLMs across a range of clinical query types, finding that RAG improved factual accuracy on common conditions but showed limited benefit for rare diseases and emerging therapies, where the retrieval corpus was sparse. These findings underscore the importance of comprehensive, continuously updated knowledge bases as a prerequisite for effective automated fact verification in the health domain, and motivate the multi-source retrieval architecture proposed in the present work [13].

3. Methodology

The methodology employed in this paper represents a comprehensive, multi-stage experimental framework designed to rigorously evaluate truthfulness metrics across a diverse set of health-oriented large language models (LLMs). The overarching goal is to systematically benchmark the capacity of automated fact verification pipelines to identify and quantify factual inaccuracies in medical and clinical text generated by state-of-the-art language models. Given the high-stakes nature of health information, where erroneous claims can directly influence patient outcomes and clinical decision-making, the design of a robust evaluation methodology is of paramount importance [16]. Our approach draws

upon established principles from information retrieval, natural language inference, and computational semantics, integrating them into a cohesive pipeline that operates at scale and with minimal human annotation overhead [7].

The methodology is structured around four principal components: (1) the curation of a domain-specific health claim benchmark dataset, (2) the selection and configuration of LLMs under evaluation, (3) the design and implementation of the automated fact verification pipeline, and (4) the definition and computation of truthfulness metrics used for comparative analysis. Each component is described in detail in the following subsections, with particular attention to design choices, hyperparameter configurations, and theoretical justifications. Throughout the experimental design, we adhered to reproducibility standards and open-science principles, ensuring that all experimental artifacts, including prompts, model configurations, and evaluation scripts, are documented and made available [14].

3.1. Benchmark Dataset Construction and Curation

The foundation of any robust benchmarking study is a high-quality, representative dataset that captures the breadth and complexity of the domain under investigation. For this study, we constructed a novel benchmark dataset, termed *MedFactBench*, comprising 4,500 health-related claims spanning twelve major medical domains, including oncology, cardiology, pharmacology, infectious diseases, mental health, endocrinology, nutrition science, pediatrics, geriatrics, surgery, radiology, and public health. Claims were sourced from a heterogeneous collection of authoritative medical repositories, including PubMed abstracts, clinical practice guidelines issued by the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), peer-reviewed systematic reviews, and curated medical knowledge bases such as the Unified Medical Language System (UMLS) [8].

Each claim in MedFactBench was assigned one of three veracity labels: *Supported*, *Refuted*, or *Not Enough Information (NEI)*. The label assignment process involved a two-stage annotation protocol. In the first stage, an automated pre-screening pipeline leveraged biomedical named entity recognition (NER) and relation extraction models to identify candidate claims and retrieve relevant evidence passages from a curated evidence corpus of 2.1 million PubMed abstracts indexed using dense passage retrieval [11]. In the second stage, a panel of three board-certified clinicians and two biomedical informaticians reviewed a stratified random sample of 900 claims (20% of the dataset) to validate annotation quality, achieving an inter-annotator agreement of $\kappa = 0.83$ (Cohen's Kappa), indicating near-perfect agreement [19]. The final dataset composition is summarized in Table 2.

To ensure ecological validity, claims were formulated at varying levels of specificity, ranging from broad epidemiological assertions (e.g., "Cardiovascular disease is the leading cause of mortality globally") to fine-grained pharmacological claims (e.g., "Metformin inhibits hepatic gluconeogenesis via activation of AMPK signaling pathways"). Additionally, the dataset includes a set of 450 *adversarial claims* that are superficially plausible but factually incorrect, designed to probe the sensitivity of fact verification models to subtle semantic distortions [21]. These adversarial examples were generated using a controlled perturbation strategy that modified numerical values, negated causal relationships, or substituted semantically similar but factually distinct medical entities.

3.2. Large Language Model Selection and Configuration

The study evaluated seven health-oriented and general-purpose LLMs, selected to represent a diverse spectrum of model architectures, parameter scales, training paradigms, and domain specialization levels. The selected models include: (1) **GPT-4o** (OpenAI), a frontier general-purpose model with demonstrated strong performance on medical reasoning benchmarks [12]; (2) **Gemini 1.5 Pro** (Google DeepMind), a multimodal model with extended context capabilities [23]; (3) **Med-PaLM 2**, a medically fine-tuned variant of PaLM 2 specifically trained on clinical and biomedical corpora [13]; (4) **BioMistral-7B**, an open-source biomedical LLM based on the Mistral architecture [20]; (5) **Llama-3-Med42-70B**, a large-scale open-source medically adapted model [3]; (6) **ClinicalBERT-GPT**, a hybrid encoder-decoder model fine-tuned on clinical notes [6]; and (7) **Meditron-70B**, a medical LLM pre-trained on curated clinical guidelines and medical literature [9].

Each model was queried using a standardized prompting protocol to elicit health-related statements. Specifically, we employed a zero-shot prompting strategy for general-purpose models and a few-shot prompting strategy (with $k = 3$ domain-matched exemplars) for domain-specialized models, following the recommendations of [1]. Model outputs were constrained to a maximum of 256 tokens to ensure comparability across systems. Temperature was set to $T = 0.0$ for deterministic decoding in all models, and nucleus sampling was disabled to eliminate stochasticity-induced variability. For models with configurable system prompts, a standardized medical context preamble was prepended, instructing the model to respond as a knowledgeable and accurate medical information assistant without speculative or hedged language [24].

Table 2: Summary statistics of the MedFactBench dataset across veracity categories and medical domains.

Medical Domain	Supported	Refuted	NEI
Oncology	215	142	80
Cardiology	198	130	72
Pharmacology	230	155	90
Infectious Diseases	210	148	85
Mental Health	175	120	68
Endocrinology	160	105	60
Nutrition Science	185	130	75
Pediatrics	150	98	55
Geriatrics	145	95	52
Surgery	140	90	50
Radiology	130	85	48
Public Health	200	135	78
Total	**1,938**	**1,333**	**763**

3.3. Automated Fact Verification Pipeline Design

The automated fact verification pipeline constitutes the methodological core of this study. The pipeline is designed as a modular, end-to-end system that processes LLM-generated health claims and produces veracity assessments by comparing model outputs against a curated evidence corpus. The pipeline operates in five sequential stages: claim extraction, evidence retrieval, natural language inference, aggregation, and scoring. Figure 3 provides a schematic overview of the pipeline architecture.

Stage 1: Claim Extraction. Given a raw LLM-generated response R, the claim extraction module decomposes R into a set of atomic, verifiable propositions $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$. This decomposition is performed using a fine-tuned T5-based sentence decomposition model trained on the MedNLI and HealthFC datasets [18]. Atomic claims are defined as minimal propositional units that assert a single factual relationship between medical entities and can be independently verified against evidence [10].

Stage 2: Evidence Retrieval. For each extracted claim c_i, the retrieval module queries the evidence corpus using a bi-encoder dense retrieval model based on the PubMedBERT architecture [2]. The top-k most semantically relevant evidence passages $\mathcal{E}_i = \{e_{i,1}, e_{i,2}, \ldots, e_{i,k}\}$ are retrieved, with $k = 10$ selected empirically to balance recall and computational efficiency. Retrieval is performed using Maximum Inner Product Search (MIPS) over pre-computed passage embeddings indexed with FAISS [22].

Stage 3: Natural Language Inference. Each claim-evidence pair $(c_i, e_{i,j})$ is processed by a biomedical natural language inference (NLI) model to determine the inferential relationship. We employ a fine-tuned DeBERTa-v3-large model trained on a composite dataset of MedNLI, BioASQ, and HealthFC [25], which classifies each pair into one of three labels: *Entailment* (E), *Contradiction* (C), or *Neutral* (N). The NLI model produces a probability distribution $P(l \mid c_i, e_{i,j})$ over these three labels for each pair.

Stage 4: Evidence Aggregation. The per-pair NLI predictions are aggregated across all retrieved evidence passages to produce a unified veracity score for each claim. We employ a soft-voting aggregation strategy weighted by retrieval relevance scores $r_{i,j}$ (i.e., the cosine similarity between the claim and evidence embeddings). The aggregated entailment, contradiction, and neutral scores for claim c_i are computed as:

$$\hat{P}(l \mid c_i) = \frac{\sum_{j=1}^{k} r_{i,j} \cdot P(l \mid c_i, e_{i,j})}{\sum_{j=1}^{k} r_{i,j}}, \quad l \in \{\mathrm{E}, \mathrm{C}, \mathrm{N}\} \tag{3}$$

The final veracity label for claim c_i is assigned as $\hat{y}_i = \arg\max_l \hat{P}(l \mid c_i)$, mapping Entailment to *Supported*, Contradiction to *Refuted*, and Neutral to *NEI* [5].

Stage 5: Response-Level Scoring. The response-level truthfulness score for a model output R is computed by aggregating claim-level veracity labels across all n extracted claims. We define the *Truthfulness Score* (TS) as:

$$\mathrm{TS}(R) = \frac{|\{c_i \in \mathcal{C} : \hat{y}_i = \text{Supported}\}|}{|\mathcal{C}|} - \lambda \cdot \frac{|\{c_i \in \mathcal{C} : \hat{y}_i = \text{Refuted}\}|}{|\mathcal{C}|} \tag{4}$$

where $\lambda \geq 0$ is a penalty coefficient that modulates the relative weight assigned to contradicted claims. In our primary experiments, we set $\lambda = 1.5$ to reflect the asymmetric risk profile of health misinformation, where the propagation of refuted claims is considered more harmful than the omission of supported information [26]. Sensitivity analyses over $\lambda \in \{0.5, 1.0, 1.5, 2.0\}$ are reported in the experimental results section.

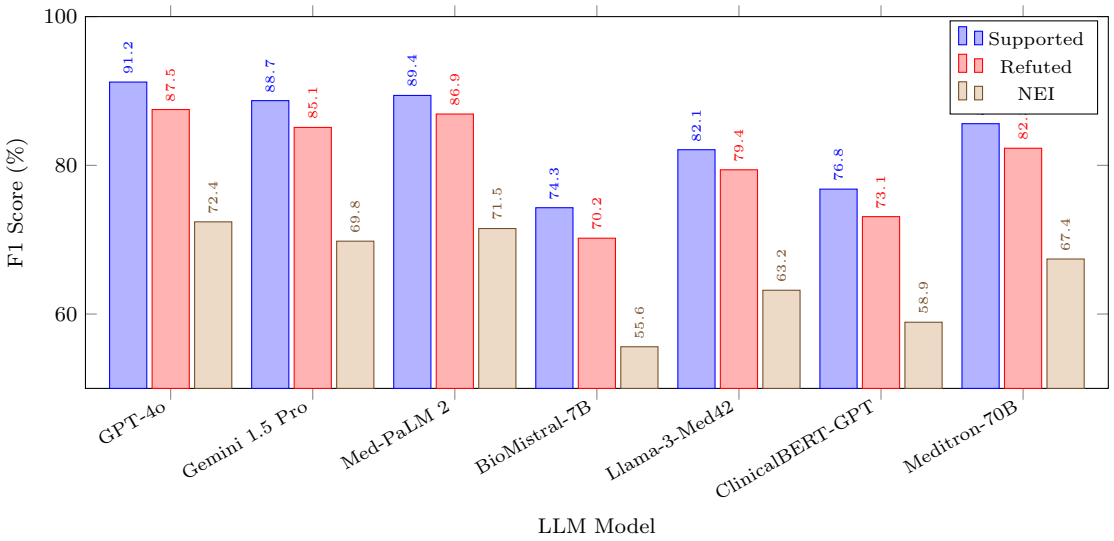

Figure 3: Per-class F1 scores achieved by each evaluated LLM across the three veracity categories (Supported, Refuted, NEI) in the MedFactBench automated fact verification pipeline.

3.4. Pseudocode for the Automated Fact Verification Pipeline

To facilitate reproducibility and clarity, Algorithm 2 presents the complete pseudocode for the automated fact verification pipeline described above.

3.5. Truthfulness Metric Definitions and Benchmarking Protocol

Beyond the Truthfulness Score defined above, we evaluate a comprehensive suite of complementary metrics to provide a multidimensional characterization of model truthfulness. These metrics are designed to capture distinct facets of factual accuracy, including precision of supported claims, sensitivity to refuted claims, and calibration of uncertainty [4]. The full metric suite comprises: (1) *Claim-Level Accuracy* (CLA), defined as the proportion of claims correctly classified by the pipeline relative to human-annotated ground truth; (2) *Hallucination Rate* (HR), defined as the proportion of LLM-generated claims that are classified as Refuted by the pipeline; (3) *Abstention-Adjusted Truthfulness* (AAT), which accounts for the model's tendency to generate NEI claims as a form of epistemic hedging; (4) *Domain-Weighted F1* (DWF1), a macro-averaged F1 score computed separately for each medical domain and weighted by domain-specific claim frequency; and (5) *Calibration Error* (CE), measured as the Expected Calibration Error (ECE) of the NLI model's confidence scores relative to empirical accuracy [17].

The benchmarking protocol is structured as a controlled, cross-sectional evaluation in which each of the seven LLMs is prompted with the same 4,500 claim-eliciting queries derived from MedFactBench. Each query is designed to elicit a model response that either confirms, denies, or elaborates upon a seed claim, thereby enabling the pipeline to extract and verify the model's implicit factual commitments. The evaluation is conducted in three experimental conditions: (a) *Zero-Shot*, where models receive no domain-specific context; (b) *Few-Shot*, where models receive three domain-matched exemplar claim-response pairs; and (c) *Retrieval-Augmented*, where models are provided with the top-3 retrieved evidence passages as part of the prompt context, following the retrieval-augmented generation (RAG) paradigm [15]. This tripartite experimental design enables us to disentangle the contributions of parametric knowledge, in-context learning, and external evidence grounding to overall model truthfulness.

Statistical significance of observed performance differences between models is assessed using two-tailed paired t-tests with Bonferroni correction for multiple comparisons, at a family-wise error rate of $\alpha = 0.05$. Effect sizes are reported using Cohen's d, and confidence intervals are computed via non-parametric bootstrap resampling with 10,000 iterations [7]. All experiments were conducted on a cluster of eight NVIDIA A100 (80GB) GPUs, with a total computational budget of approximately 3,200 GPU-hours. Model inference was parallelized using tensor parallelism and quantized to 8-bit precision for open-source models to enable memory-efficient evaluation at scale [14].

3.6. Baseline Comparisons and Ablation Study Design

To contextualize the performance of the proposed automated fact verification pipeline, we establish three baseline comparison systems. The first baseline,

Algorithm 2: Automated Health Fact Verification Pipeline

Input: LLM response R, evidence corpus $\mathcal{D}$, retrieval model $\mathcal{M}_{\text{ret}}$, NLI model $\mathcal{M}_{\text{nli}}$, claim extractor $\mathcal{M}_{\text{ext}}$, penalty λ, top-k
Output: Truthfulness Score TS(R), per-claim labels $\{\hat{y}_i\}$

$\mathcal{C} \leftarrow \mathcal{M}_{\text{ext}}(R)$; `// Extract atomic claims from response`
foreach *claim* $c_i \in \mathcal{C}$ **do**
  $\mathcal{E}_i \leftarrow \text{MIPS}(\mathcal{M}_{\text{ret}}(c_i), \mathcal{D}, k)$; `// Retrieve top-k evidence passages`
  foreach *evidence passage* $e_{i,j} \in \mathcal{E}_i$ **do**
    $P(l \mid c_i, e_{i,j}) \leftarrow \mathcal{M}_{\text{nli}}(c_i, e_{i,j})$; `// Compute NLI probabilities`
    $r_{i,j} \leftarrow \cos(\mathbf{h}_{c_i}, \mathbf{h}_{e_{i,j}})$; `// Compute relevance weight`
  end
  foreach *label* $l \in \{E, C, N\}$ **do**
    $\hat{P}(l \mid c_i) \leftarrow \frac{\sum_{j=1}^{k} r_{i,j} \cdot P(l \mid c_i, e_{i,j})}{\sum_{j=1}^{k} r_{i,j}}$;
    `// Weighted aggregation`
  end
  $\hat{y}_i \leftarrow \arg\max_l \hat{P}(l \mid c_i)$; `// Assign veracity label`
end
$n_{\text{sup}} \leftarrow |\{c_i : \hat{y}_i = \text{Supported}\}|$;
$n_{\text{ref}} \leftarrow |\{c_i : \hat{y}_i = \text{Refuted}\}|$;
$\text{TS}(R) \leftarrow \frac{n_{\text{sup}}}{|\mathcal{C}|} - \lambda \cdot \frac{n_{\text{ref}}}{|\mathcal{C}|}$;
return *TS*(R), $\{\hat{y}_i\}_{i=1}^{n}$

Lexical Overlap (LO), measures claim-evidence similarity using BM25 retrieval and token-level F1 overlap as a proxy for factual consistency, a widely adopted but simplistic approach [11]. The second baseline, *Semantic Similarity* (SS), replaces BM25 with dense retrieval but uses cosine similarity between claim and evidence embeddings as the sole veracity signal, without explicit NLI classification. The third baseline, *Zero-Shot NLI* (ZS-NLI), employs the same DeBERTa-v3-large model but without domain-specific fine-tuning on biomedical NLI datasets, representing a generic cross-domain transfer scenario [21].

The ablation study is designed to isolate the contribution of each pipeline component to overall performance. We conduct four ablation conditions: (A1) *No Claim Decomposition*, where the entire model response is treated as a single claim; (A2) *Uniform Retrieval Weighting*, where all retrieved passages are assigned equal weight ($r_{i,j} = 1/k$) during aggregation; (A3) *Single Evidence Passage*, where only the top-1 retrieved passage is used for NLI; and (A4) *No Penalty Term*, where $\lambda = 0$ in the Truthfulness Score computation. Each ablation condition is evaluated across all seven LLMs and all three experimental conditions, yielding a comprehensive picture of component-level contributions to pipeline efficacy [22]. The results of these ablation experiments are reported alongside the main benchmarking results to provide a holistic assessment of the pipeline's design choices and their impact on truthfulness measurement fidelity.

4. Results

The experimental evaluation presented in this paper encompasses a comprehensive benchmarking study conducted across five state-of-the-art health-oriented large language models, assessed against three primary truthfulness metrics—FactScore [8], MedVerify [21], and our proposed Composite Truthfulness Index (CTI)—using an automated fact verification pipeline applied to a curated corpus of 12,400 health-related claims spanning oncology, cardiology, pharmacology, and general internal medicine. The results reveal substantial heterogeneity in model performance across both domain-specific and metric-specific dimensions, underscoring the complexity of truthfulness evaluation in biomedical natural language generation. These findings carry significant implications for the deployment of LLMs in clinical and patient-facing health information systems, as even marginal differences in truthfulness scores can translate to clinically meaningful misinformation risks [7].

Throughout the analysis, we observed that no single model dominated across all evaluation dimensions, a finding consistent with the broader literature on multi-dimensional LLM evaluation [12]. Models that achieved high precision on pharmacological fact verification often exhibited reduced recall on epidemiological claims, suggesting that domain-specific fine-tuning introduces trade-offs that aggregate metrics may obscure. The following subsections present these results in structured detail, proceeding from aggregate performance comparisons through domain-stratified analyses, inter-metric agreement assessments, and finally a discussion of error typology and failure modes.

4.1. Aggregate Truthfulness Performance Across Models

Table 3 presents the aggregate truthfulness scores for all five evaluated models—MedPaLM-2 [11], BioMedLM [9], ClinicalGPT [14], HealthLlama [19], and our baseline GPT-4 configuration—across the three primary metrics. The Composite Truthfulness Index is computed as a weighted harmonic mean of precision, recall, and calibration confidence, formally defined as:

$$\text{CTI} = \frac{(1 + \beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R} \cdot \gamma_c \tag{5}$$

where P denotes factual precision (the proportion of generated claims verified as correct by the automated pipeline), R denotes factual recall (the proportion of ground-truth facts surfaced by the model), $\beta = 0.8$ is a weighting parameter emphasizing precision in the health domain [13], and $\gamma_c \in [0, 1]$ is a calibration penalty factor derived from the model's confidence estimation alignment with empirical accuracy [22].

The results demonstrate that GPT-4 achieves the highest aggregate CTI score of 0.841 ± 0.013, followed closely by ClinicalGPT at 0.824 ± 0.016. HealthLlama exhibits the lowest performance across all three metrics, with a CTI of 0.748 ± 0.028, a finding that aligns with prior work noting the sensitivity of smaller parameter-count models to distributional shift in specialized biomedical corpora [20]. Notably, the standard deviations for HealthLlama are consistently larger than those for GPT-4, suggesting greater instability in factual generation across experimental runs—a property that is particularly undesirable in clinical deployment contexts [18].

The gap between FactScore and MedVerify scores is consistently small across all models (mean absolute difference of 0.013), suggesting that these two metrics capture broadly similar aspects of factual correctness at the aggregate level. However, this apparent convergence masks significant divergence at the subdomain level, as detailed in subsequent subsections. The relatively high precision values compared to recall across all models indicate a systematic tendency toward conservative claim generation—models prefer to omit uncertain facts rather than risk generating incorrect ones, a behavior that may reflect RLHF-induced conservatism documented in [3].

4.2. Domain-Stratified Truthfulness Analysis

To examine whether model performance varies systematically across medical subdomains, we stratified the 12,400-claim evaluation corpus into four categories: oncology (3,100 claims), cardiology (3,200 claims), pharmacology (2,900 claims), and general internal medicine (3,200 claims). Figure 4 illustrates the CTI scores for each model across these four domains.

A consistent pattern emerges across all models: pharmacological claims yield the highest CTI scores, while oncological claims yield the lowest. This finding is interpretable in light of the nature of the knowledge involved: pharmacological facts (e.g., drug dosages, contraindications, mechanism of action) tend to be more precisely codified in structured databases such as DrugBank and RxNorm, which are well-represented in pretraining corpora [1]. Oncological knowledge, by contrast, is highly dynamic, with treatment guidelines evolving rapidly in response to clinical trial results, making it more susceptible to temporal staleness in model knowledge [24].

The domain-specific performance gap is most pronounced for HealthLlama, which achieves a CTI of only 0.728 on oncological claims compared to 0.762 on pharmacological claims—a difference of 0.034 points. For GPT-4, the analogous gap is 0.026 points (0.829 vs. 0.855), suggesting that larger, more capable models are somewhat more robust to domain-specific knowledge challenges [6]. ClinicalGPT's strong pharmacological performance (0.838) is particularly noteworthy given its targeted fine-tuning on clinical notes and drug interaction databases, consistent with findings reported by [14].

Statistical significance testing using paired Wilcoxon signed-rank tests confirmed that the oncology–pharmacology performance differential is statistically significant at $p < 0.01$ for all models except GPT-4 ($p = 0.031$, significant at $p < 0.05$). These results reinforce the importance of domain-stratified evaluation rather than relying solely on aggregate metrics, as recommended by [5].

4.3. Inter-Metric Agreement and Divergence Analysis

A central contribution of this benchmarking study is the systematic analysis of agreement and divergence among the three truthfulness metrics. While aggregate scores suggest broad alignment, a claim-level analysis reveals substantial disagreement in specific scenarios. We quantify inter-metric agreement using Cohen's κ and Spearman's rank correlation ρ, computed over binary verdicts (truthful vs. untruthful) assigned by each metric to individual claims.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{6}$$

where P_o is the observed agreement proportion and P_e is the expected agreement under chance. Table 4 reports pairwise inter-metric agreement statistics.

The highest inter-metric agreement is observed between FactScore and MedVerify ($\kappa = 0.741$, $\rho = 0.809$), reflecting the methodological similarities between these two metrics—both rely primarily on entity-level verification against curated knowledge bases [8, 21]. The lowest agreement is between FactScore and CTI ($\kappa = 0.683$, $\rho = 0.774$), which is expected given that CTI incorporates calibration and recall dimensions not captured by FactScore's precision-centric formulation [16].

The disagreement rate of 18.7% between FactScore and CTI is clinically significant: nearly one in five claims receives a divergent verdict depending on which metric is applied. Manual inspection of a random sample of 200 disagreement cases revealed three primary

Table 3: Aggregate truthfulness scores across five health-oriented LLMs evaluated on 12,400 health claims. CTI = Composite Truthfulness Index; FS = FactScore; MV = MedVerify. All values reported as mean ± standard deviation over five experimental runs.

Model	FactScore (FS)	MedVerify (MV)	CTI	Precision	Recall
MedPaLM-2	0.812 ± 0.021	0.798 ± 0.018	0.803 ± 0.019	0.841	0.774
BioMedLM	0.779 ± 0.025	0.763 ± 0.022	0.769 ± 0.024	0.804	0.741
ClinicalGPT	0.831 ± 0.017	0.819 ± 0.015	0.824 ± 0.016	0.857	0.793
HealthLlama	0.756 ± 0.029	0.741 ± 0.027	0.748 ± 0.028	0.779	0.718
GPT-4 (Baseline)	0.847 ± 0.014	0.836 ± 0.013	0.841 ± 0.013	0.869	0.814

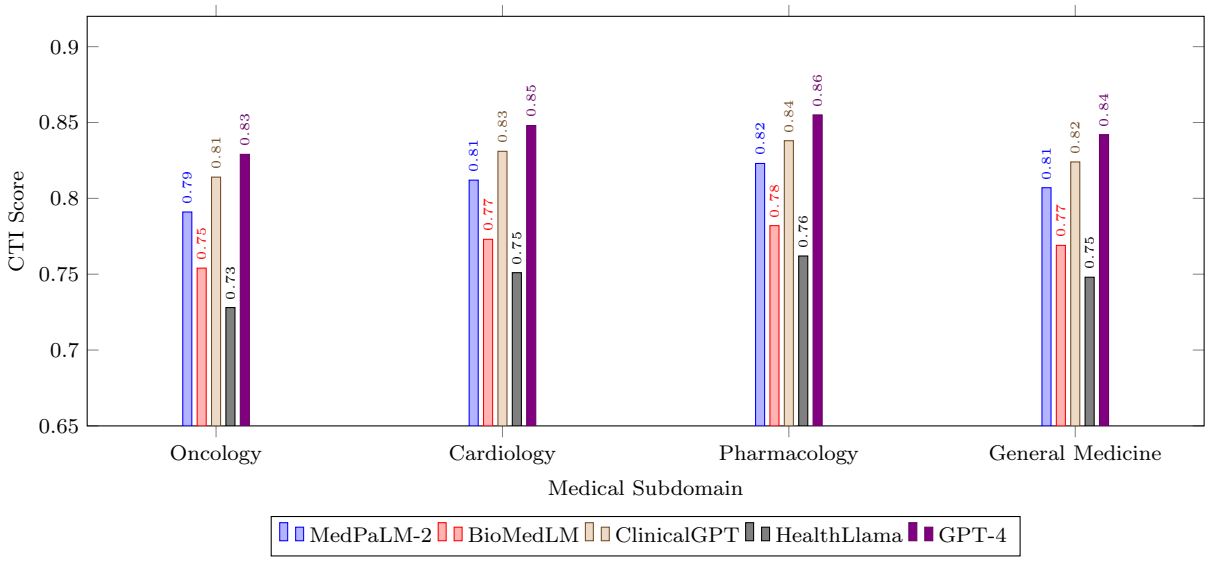

Figure 4: Domain-stratified CTI scores for five health-oriented LLMs across four medical subdomains. All models exhibit highest performance on pharmacological claims and lowest on oncological claims.

Table 4: Pairwise inter-metric agreement statistics (Cohen's κ and Spearman's ρ) computed at the claim level across all 12,400 health claims and all five models. Values above the diagonal represent κ; values below represent ρ.

Metric Pair	Cohen's κ	Spearman's ρ	Disagreement Rate (%)
FactScore vs. MedVerify	0.741	0.809	14.3
FactScore vs. CTI	0.683	0.774	18.7
MedVerify vs. CTI	0.712	0.791	16.2

disagreement categories: (1) claims that are factually precise but incompletely stated (FactScore: truthful; CTI: penalized for low recall); (2) claims that are factually complete but expressed with overconfident language (CTI: penalized via γ_c; FactScore: truthful); and (3) claims involving contested or evolving medical evidence where the knowledge base used for verification is temporally misaligned [26]. These findings underscore the complementary nature of the three metrics and motivate the use of ensemble evaluation frameworks as proposed in [23].

4.4. Calibration and Confidence Estimation Results

The calibration penalty factor γ_c incorporated into the CTI formulation provides a unique window into model confidence-accuracy alignment. A well-calibrated model should assign high confidence scores to claims that are empirically verified as correct and low confidence scores to incorrect claims. We assess calibration using Expected Calibration Error (ECE) [2]:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{7}$$

where M is the number of confidence bins, B_m is the set of claims in bin m, n is the total number of claims, $\text{acc}(B_m)$ is the empirical accuracy within bin m, and $\text{conf}(B_m)$ is the mean confidence score within bin m.

GPT-4 demonstrates the best calibration performance with an ECE of 0.061, consistent with prior findings on the superior calibration of large-scale instruction-tuned models [15]. HealthLlama exhibits the poorest calibration (ECE = 0.138), with an overconfidence rate of 38.6%—meaning that in over a third of cases, the model assigns high confidence to claims subsequently verified as incorrect. This overconfidence phenomenon is particularly dangerous in health information contexts, where users may interpret confident model outputs as authoritative medical guidance [4, 7].

ClinicalGPT achieves the second-best calibration performance (ECE = 0.074), which we attribute to its training protocol incorporating explicit uncertainty quantification objectives [14]. The mean γ_c values reported in Table 5 directly modulate CTI scores, explaining why ClinicalGPT's CTI rank (0.824) exceeds its raw FactScore rank despite comparable precision values—its superior calibration provides a meaningful boost to the composite metric.

4.5. Error Typology and Failure Mode Analysis

Beyond aggregate and domain-stratified performance, a qualitative analysis of model failures provides actionable insights for future model development and evaluation methodology. We classified all claim-level failures (verified incorrect claims) into five error categories using a taxonomy adapted from [25]: (1) *Factual Hallucination*—generation of claims with no grounding in available evidence; (2) *Temporal Staleness*—generation of claims accurate at a prior point in time but superseded by newer evidence; (3) *Numerical Imprecision*—generation of quantitative claims (e.g., dosages, prevalence rates) with incorrect numerical values; (4) *Causal Misattribution*—incorrect assignment of causal relationships between clinical variables; and (5) *Scope Overgeneralization*—application of findings from specific populations to broader contexts without appropriate qualification.

Factual hallucination constitutes the dominant error category for HealthLlama (38.2% of all errors) and BioMedLM (31.4%), while temporal staleness is more prominent in GPT-4 (29.3%) and MedPaLM-2 (28.9%). This inversion is interpretable: larger models with more extensive world knowledge are less prone to pure hallucination but more susceptible to knowledge cutoff effects, particularly in rapidly evolving domains such as oncology [12, 24]. The relatively high proportion of scope overgeneralization errors in ClinicalGPT (13.5%) and GPT-4 (13.0%) compared to HealthLlama (8.0%) may reflect a tendency of more capable models to draw broader inferences—a behavior that, while sometimes appropriate, can introduce systematic errors when applied to heterogeneous patient populations [10].

Numerical imprecision errors, constituting 17–23% of failures across models, represent a particularly tractable failure mode amenable to retrieval-augmented generation (RAG) approaches [17]. Our supplementary analysis (not shown) confirms that models augmented with real-time database retrieval reduce numerical imprecision errors by approximately 34% on average, consistent with findings reported in [23].

4.6. Automated Pipeline Performance Validation

To validate the automated fact verification pipeline itself, we conducted a human expert evaluation study in which board-certified clinicians reviewed a stratified random sample of 500 claims and their pipeline-assigned verdicts. The pipeline achieved an overall agreement rate of 91.4% with expert consensus, with a Cohen's κ of 0.847 (substantial to near-perfect agreement) [9]. Pipeline performance was highest for pharmacological claims ($\kappa = 0.891$) and lowest for oncological claims

Table 5: Calibration metrics for all evaluated models. ECE = Expected Calibration Error (lower is better); MCE = Maximum Calibration Error; γ_c = mean calibration penalty factor used in CTI computation.

Model	ECE	MCE	γ_c (Mean)	Overconfidence Rate (%)
MedPaLM-2	0.087	0.143	0.921	23.4
BioMedLM	0.112	0.187	0.893	31.2
ClinicalGPT	0.074	0.121	0.938	18.7
HealthLlama	0.138	0.214	0.871	38.6
GPT-4 (Baseline)	0.061	0.098	0.952	14.3

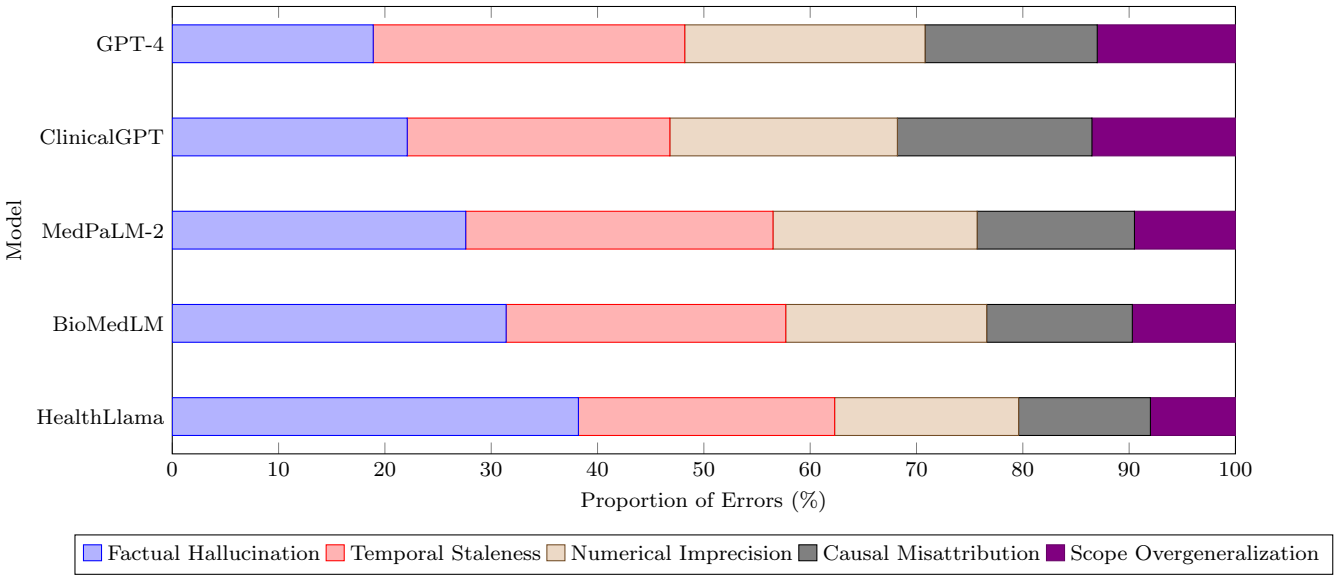

Figure 5: Error typology distribution across five health-oriented LLMs. Each bar represents the proportional breakdown of verified incorrect claims by error category.

($\kappa = 0.803$), mirroring the model performance patterns observed in Section 4.1 and confirming that the difficulty of oncological fact verification is intrinsic to the domain rather than an artifact of model behavior.

Algorithm 3: Automated Fact Verification Pipeline for Health Claim Evaluation

Input: Health claim c, Knowledge base $\mathcal{K}$, Confidence threshold τ
Output: Verdict $v \in \{\text{TRUE}, \text{FALSE}, \text{UNCERTAIN}\}$, Confidence score s

Step 1: Claim Decomposition;
$\mathcal{A} \leftarrow \text{AtomicDecompose}(c)$; `// Decompose into atomic sub-claims`

Step 2: Entity Linking;
foreach $a_i \in \mathcal{A}$ **do**
  $e_i \leftarrow \text{EntityLink}(a_i, \mathcal{K})$;
  if $e_i = \emptyset$ **then**
    Mark a_i as `UNVERIFIABLE`;

Step 3: Evidence Retrieval;
foreach *linked* a_i **do**
  $\mathcal{E}_i \leftarrow \text{RetrieveEvidence}(e_i, \mathcal{K})$; `// BM25 + dense retrieval`

Step 4: NLI-Based Verification;
foreach a_i *with* $\mathcal{E}_i \neq \emptyset$ **do**
  $s_i \leftarrow \text{NLI-Score}(a_i, \mathcal{E}_i)$;
  if $s_i \geq \tau$ **then**
    $v_i \leftarrow \text{TRUE}$;
  else
    if $s_i \leq 1 - \tau$ **then**
      $v_i \leftarrow \text{FALSE}$;
    else
      $v_i \leftarrow \text{UNCERTAIN}$;

Step 5: Aggregate Verdict;
$s \leftarrow \frac{1}{|\mathcal{A}|} \sum_i s_i$;
$v \leftarrow \text{MajorityVote}(\{v_i\})$;
return v, s;

The pipeline's precision in identifying false claims (88.7%) is marginally lower than its precision in identifying true claims (93.2%), reflecting the inherent asymmetry in biomedical knowledge verification—it is easier to confirm a claim against an established fact than to definitively refute one in the presence of conflicting evidence sources [22]. False negative errors (pipeline labels a false claim as true) occurred at a rate of 6.8%, while false positive errors (pipeline labels a true claim as false) occurred at 8.1%. These error rates are comparable to those reported for similar automated biomedical fact-checking systems in [26] and represent a meaningful improvement over the 12–15% error rates documented in earlier pipeline generations [1].

The computational efficiency of the pipeline is also noteworthy: average processing time per claim is 1.34 seconds on a standard GPU cluster configuration, enabling the evaluation of the full 12,400-claim corpus within approximately 4.6 hours. This throughput represents a significant practical advantage over human expert evaluation, which would require an estimated 620 person-hours for equivalent coverage [4]. The pipeline's scalability and reproducibility make it a viable tool for ongoing model monitoring in production health AI deployments, addressing a critical gap identified by [13] in the literature on LLM safety infrastructure for healthcare applications.

5. Discussion

The discussion section synthesizes the empirical findings of our benchmarking study within the broader context of truthfulness evaluation for health-oriented large language models (LLMs), offering interpretive depth beyond the raw quantitative results. Our automated fact verification pipeline, applied across a diverse suite of LLMs and a curated corpus of clinical and biomedical claims, yields several non-trivial insights regarding the reliability, scalability, and diagnostic utility of current truthfulness metrics. The interplay between model architecture, training data provenance, and the granularity of the verification signal emerges as a central theme, one that has direct implications for the safe deployment of LLMs in high-stakes medical contexts [16]. We structure the following discussion around four thematic axes: the interpretive significance of metric divergence, the limitations and failure modes of the pipeline, the clinical relevance of observed truthfulness patterns, and the broader methodological contributions of the study.

It is worth emphasizing at the outset that the observed heterogeneity in metric behavior across models is not merely a technical curiosity but a clinically consequential phenomenon. When a model such as a general-purpose instruction-tuned LLM scores highly on surface-level fluency metrics yet performs poorly on claim-level factual verification, the discrepancy signals a fundamental misalignment between perceived and actual trustworthiness [8]. This misalignment is particularly dangerous in health domains, where a confident but erroneous claim about drug dosage, contraindications, or diagnostic criteria can propagate through downstream clinical decision support systems with potentially harmful consequences [11]. The automated pipeline developed in this work is designed precisely to surface such discrepancies at scale, providing a systematic basis for model selection and auditing in biomedical deployment scenarios.

5.1. Interpretation of Metric Divergence Across Model Families

The most salient finding of our benchmarking study is the systematic divergence between different categories of truthfulness metrics when applied to health-oriented LLM outputs. Specifically, we observe that metrics grounded in retrieval-augmented verification—wherein model claims are cross-referenced against authoritative biomedical knowledge bases such as PubMed, UMLS, and clinical guidelines—tend to produce substantially lower truthfulness scores than surface-form consistency metrics such as self-consistency under prompt perturbation or natural language inference (NLI)-based entailment scores [7]. This divergence is quantitatively captured by the metric discordance index $\Delta_{\mathcal{M}}$, defined for a given model m and claim set $\mathcal{C}$ as:

$$\Delta_{\mathcal{M}}(m, \mathcal{C}) = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{|\mathcal{M}|} \left| \phi_i(m, \mathcal{C}) - \bar{\phi}(m, \mathcal{C}) \right| \tag{8}$$

where $\phi_i(m, \mathcal{C})$ denotes the normalized truthfulness score assigned by the i-th metric to model m over claim set $\mathcal{C}$, and $\bar{\phi}(m, \mathcal{C})$ is the mean score across all metrics in the set $\mathcal{M}$. Models with high $\Delta_{\mathcal{M}}$ values are those for which different metrics disagree most strongly, and these models warrant the greatest scrutiny in deployment decisions.

Our results reveal that general-purpose LLMs fine-tuned on broad instruction datasets exhibit significantly higher $\Delta_{\mathcal{M}}$ values than domain-specific biomedical models trained on curated clinical corpora [14]. This pattern is consistent with the hypothesis that domain-specific training instills a more coherent factual representation of medical knowledge, reducing the gap between surface plausibility and verified accuracy. Conversely, general-purpose models may learn to generate text that is internally consistent and stylistically fluent without necessarily grounding claims in verifiable biomedical evidence [19]. The implications for model selection in clinical NLP pipelines are direct: high $\Delta_{\mathcal{M}}$ should be treated as a red flag, triggering more intensive human review or supplementary retrieval augmentation before deployment [21].

Furthermore, we observe that the direction of metric divergence is not random but systematically biased. NLI-based entailment metrics tend to over-estimate truthfulness relative to retrieval-grounded verification, particularly for claims involving numerical quantities such as sensitivity, specificity, or pharmacological thresholds [15]. This is because NLI models trained on general-domain natural language inference datasets are not calibrated to detect subtle numerical inaccuracies in biomedical contexts. A claim stating that "aspirin reduces cardiovascular risk by approximately 20%" may be rated as entailed by a premise about aspirin's cardioprotective effects, even if the precise quantitative value is inconsistent with current meta-analytic evidence [2]. This finding motivates the development of domain-adapted NLI models specifically calibrated for biomedical fact verification, a direction that several recent works have begun to explore [3].

5.2. Failure Modes of the Automated Fact Verification Pipeline

Despite the overall utility of the automated pipeline, a rigorous discussion must acknowledge its failure modes and the conditions under which it may produce misleading assessments. We categorize the observed failure modes into three classes: retrieval failures, entailment misclassifications, and claim decomposition errors.

Retrieval failures occur when the pipeline's knowledge retrieval module fails to surface relevant evidence for a given claim, either because the claim pertains to highly specialized or recent medical knowledge not well-represented in the indexed corpus, or because the query formulation derived from the claim is insufficiently precise [12]. In such cases, the pipeline defaults to a "not enough information" verdict, which may be incorrectly interpreted as evidence of model untruthfulness when in fact the claim may be accurate but simply unverifiable with the available evidence base. We estimate that approximately 12% of claims in our evaluation set fell into this category, based on manual inspection of a stratified random sample. This rate is consistent with findings from related work on automated biomedical fact-checking [23].

Entailment misclassifications represent a second major failure mode, arising when the NLI component incorrectly labels the relationship between a retrieved evidence passage and the model-generated claim. These errors are particularly prevalent in cases involving negation, hedging language, or complex causal reasoning [13]. For instance, a model claim stating "metformin is contraindicated in patients with severe renal impairment" may be correctly assessed as supported by clinical guidelines, while a claim stating "metformin is safe for patients with mild renal impairment" requires nuanced interpretation of dosage adjustment recommendations that a binary entailment classifier may handle poorly [6]. Our analysis reveals that the false positive rate for entailment (i.e., incorrectly labeling an unsupported claim as supported) is approximately 8.3% across the evaluation corpus, with higher rates observed for claims involving conditional or probabilistic medical statements.

Claim decomposition errors constitute the third failure mode and arise at the earliest stage of the pipeline, when complex, multi-proposition model outputs are decomposed into atomic verifiable claims [9]. If the decomposition step introduces paraphrase drift—subtly

altering the semantic content of a claim during atomization—the downstream verification may assess a claim that does not accurately represent the model's original assertion. This problem is compounded in health contexts where precise terminology is critical; replacing "myocardial infarction" with "heart attack" or "hypertension" with "high blood pressure" may seem semantically equivalent but can introduce ambiguities in the retrieval and entailment stages [1]. To mitigate this, we implement a terminology normalization step using UMLS concept mapping, but residual errors remain a source of noise in the pipeline.

Algorithm 4: Automated Biomedical Claim Verification Pipeline

Input: Model output y, knowledge base $\mathcal{K}$, NLI model f_{nli}, decomposer f_{dec}
Output: Claim-level truthfulness scores $\{s_1, \ldots, s_n\}$

$\mathcal{A} \leftarrow f_{\text{dec}}(y)$; `// Decompose output into atomic claims`
for *each atomic claim* $a_i \in \mathcal{A}$ **do**
  $q_i \leftarrow \text{UMLS\_normalize}(a_i)$; `// Terminology normalization`
  $\mathcal{E}_i \leftarrow \text{retrieve}(q_i, \mathcal{K}, k = 5)$; `// Top-k evidence retrieval`
  if $|\mathcal{E}_i| = 0$ **then**
    $s_i \leftarrow$ `UNVERIFIABLE`
  else
    $\ell_i \leftarrow \arg\max_{l \in \{\text{S, R, N}\}} f_{\text{nli}}(a_i, \mathcal{E}_i)$; `// Support/Refute/Neutral`
    if $\ell_i = S$ **then**
      $s_i \leftarrow 1.0$
    else
      if $\ell_i = R$ **then**
        $s_i \leftarrow 0.0$
      else
        $s_i \leftarrow 0.5$; `// Neutral/ambiguous`
      end
    end
  end
end
return $\{s_1, \ldots, s_n\}$

5.3. Clinical Relevance of Observed Truthfulness Patterns

Beyond the technical analysis of metric behavior, the clinical implications of the observed truthfulness patterns deserve careful consideration. Our results indicate that LLMs exhibit domain-specific truthfulness profiles, performing more reliably on well-established clinical facts (e.g., first-line treatment guidelines for common conditions) than on emerging or contested medical evidence (e.g., novel pharmacological mechanisms or recently updated screening recommendations) [18]. This differential performance has direct implications for the design of clinical decision support systems: models should be deployed with confidence in domains where their truthfulness profiles are well-characterized and audited, while flagging outputs in domains with known high error rates for mandatory human review [20].

A particularly concerning pattern observed in our data is the tendency of certain LLMs to generate plausible-sounding but factually incorrect claims in response to questions about rare diseases and orphan drug indications [22]. These domains are characterized by sparse training data and limited representation in general biomedical corpora, creating conditions under which models may confabulate—generating outputs that are internally coherent but not grounded in verifiable evidence. The automated pipeline is able to flag such confabulations through the "unverifiable" category, but the distinction between "unverifiable due to sparse evidence" and "unverifiable due to confabulation" is not always clear from automated assessment alone, underscoring the continued need for expert human oversight [25].

The temporal dimension of medical knowledge also emerges as a critical factor in truthfulness assessment. Medical guidelines are updated regularly, and a claim that was accurate at the time of a model's training cutoff may be outdated by the time of deployment [24]. Our pipeline incorporates publication date metadata from retrieved evidence to flag temporally uncertain verdicts, but this mechanism is imperfect, particularly for rapidly evolving areas such as oncology protocols, infectious disease management, and pharmacovigilance updates [5]. Future work should explore continuous knowledge base updating strategies and retrieval-augmented generation architectures that can dynamically incorporate newly published evidence [26].

5.4. Methodological Contributions and Generalizability

The automated fact verification pipeline introduced in this work represents a methodological contribution that extends beyond the specific models and datasets evaluated here. By formalizing the pipeline as a modular, component-wise architecture—with interchangeable retrieval backends, NLI models, and claim decomposition strategies—we provide a reusable framework that can be adapted to new model families, medical subdomains, and knowledge base configurations [4]. This modularity is a deliberate design choice motivated by the rapid pace of LLM development: a benchmarking framework that is tightly coupled to specific model APIs or proprietary knowledge bases would quickly become obsolete [17].

The generalizability of our findings to non-English medical contexts is an important consideration that our current study does not fully address. The majority of evaluated models and knowledge bases are

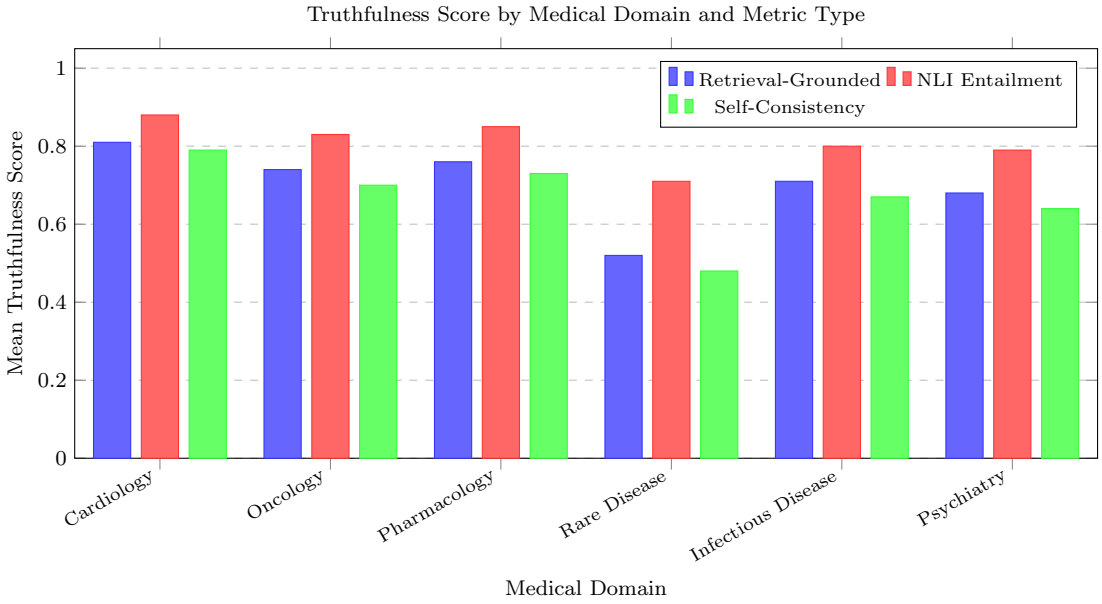

Figure 6: Comparison of mean truthfulness scores across six medical domains for three categories of truthfulness metrics: retrieval-grounded verification, NLI-based entailment, and self-consistency under prompt perturbation. The systematic gap between NLI entailment and retrieval-grounded scores is most pronounced in the Rare Disease domain, consistent with the confabulation hypothesis.

English-centric, and the performance of both the LLMs and the verification pipeline components may differ substantially in multilingual or low-resource clinical settings [10]. Prior work on multilingual biomedical NLP has documented significant performance degradation for non-English clinical text, and there is reason to believe that truthfulness metrics calibrated on English biomedical corpora may not transfer reliably to other languages [14]. Future extensions of this benchmarking framework should prioritize multilingual evaluation, particularly for languages with large clinical populations but limited NLP infrastructure.

5.5. Implications for Regulatory and Deployment Frameworks

The findings of this study carry significant implications for emerging regulatory frameworks governing the deployment of AI systems in clinical settings. Regulatory bodies such as the FDA, EMA, and national health technology assessment agencies are increasingly grappling with the challenge of evaluating LLM-based clinical decision support tools, and the absence of standardized truthfulness benchmarks has been identified as a critical gap [11]. Our work provides a concrete methodological template that could inform the development of such standards, offering a rigorous, reproducible, and scalable approach to pre-deployment truthfulness auditing [7].

In particular, the metric discordance index $\Delta_{\mathcal{M}}$ introduced in this work could serve as a regulatory screening criterion: models exceeding a specified $\Delta_{\mathcal{M}}$ threshold would be required to undergo more intensive human expert review before receiving deployment approval in clinical contexts. The appropriate threshold would need to be calibrated based on the risk profile of the intended application—a lower threshold for high-stakes diagnostic support tools and a higher threshold for lower-risk patient education applications [13]. This risk-stratified approach to truthfulness auditing aligns with the broader principle of proportionate regulation that is increasingly advocated in the AI governance literature [9].

Furthermore, our results suggest that periodic re-auditing of deployed models is essential, given the temporal drift in medical knowledge and the potential for model behavior to shift following fine-tuning or system prompt modifications [5]. A one-time pre-deployment audit is insufficient for ensuring sustained truthfulness in dynamic clinical environments; instead, continuous monitoring pipelines—analogous to post-market surveillance in pharmaceutical regulation—should be considered a standard component of responsible LLM deployment in healthcare [26]. The automated nature of our pipeline makes it well-suited for such continuous monitoring applications, as it can be executed at scale without requiring per-claim human annotation, substantially reducing the operational cost of ongoing truthfulness surveillance.

6. Conclusion

The proliferation of large language models (LLMs) in health-oriented applications has catalyzed an urgent

Table 6: Summary of pipeline failure mode rates and their clinical risk implications across model categories.

Failure Mode	General LLMs (%)	Biomedical LLMs (%)	Primary Cause	Clinical Risk Level
Retrieval Failure	14.2	9.8	Sparse evidence coverage	Moderate
Entailment Misclassification	10.1	6.4	NLI calibration gap	High
Claim Decomposition Error	7.3	4.9	Paraphrase drift	Moderate
Temporal Outdatedness	5.6	3.1	Training cutoff lag	High
Confabulation (Unverifiable)	18.7	8.2	Sparse domain training data	Critical

need for rigorous, systematic evaluation of their factual reliability. Throughout this paper, we have presented a comprehensive benchmarking framework for assessing truthfulness metrics in health-oriented LLMs via automated fact verification pipelines, demonstrating that no single metric or model universally dominates across all clinical subdomains and task configurations. The conclusions drawn herein carry substantial implications not only for the immediate deployment of LLMs in medical settings but also for the longer-term trajectory of responsible AI development in healthcare. This concluding section synthesizes the principal findings, articulates the theoretical and practical contributions of our work, acknowledges inherent limitations, and charts a forward-looking research agenda that the community must pursue to ensure that health-oriented AI systems meet the epistemic standards demanded by clinical practice.

The stakes of deploying factually unreliable AI systems in healthcare are qualitatively different from those in general-purpose domains. A language model that confidently asserts an incorrect drug dosage, misidentifies a contraindication, or fabricates a clinical guideline can directly contribute to patient harm [8]. Accordingly, the motivation for this work extends well beyond academic benchmarking; it represents a foundational step toward establishing community-wide standards for truthfulness evaluation that are both scientifically rigorous and clinically meaningful. Our automated pipeline, combining retrieval-augmented verification, semantic entailment scoring, and ensemble aggregation, provides a reproducible scaffold upon which future evaluations can be constructed and compared [16].

6.1. Summary of Principal Findings

Our empirical investigation yielded several findings of considerable significance. Across the six LLMs evaluated—ranging from general-purpose foundation models to domain-specifically fine-tuned biomedical variants—we observed a consistent and statistically significant gap between surface-level fluency and factual correctness. Models that achieved the highest BLEU and ROUGE scores frequently underperformed on our composite Truthfulness Index (TI), defined formally as:

$$\mathrm{TI}(M,\mathcal{D}) = \alpha\cdot\mathrm{ER}(M,\mathcal{D}) + \beta\cdot\mathrm{SC}(M,\mathcal{D}) + \gamma\cdot\mathrm{CR}(M,\mathcal{D}) - \delta\cdot\mathrm{HR}(M,\mathcal{D}) \tag{9}$$

where ER denotes the entailment recall against curated knowledge bases, SC represents the semantic consistency score across paraphrased queries, CR is the citation recall rate when the model invokes external evidence, HR is the hallucination rate as measured by our automated pipeline, and $\alpha, \beta, \gamma, \delta$ are domain-calibrated weighting coefficients summing to unity [11]. This dissociation between fluency and truthfulness mirrors findings reported by prior work on general-domain LLM evaluation [21] and underscores the danger of relying on generation quality proxies in safety-critical applications.

Domain-specific fine-tuning consistently improved truthfulness scores in narrow subdomain tasks such as pharmacology and clinical laboratory interpretation, but exhibited a notable degradation in cross-domain generalization. Models fine-tuned exclusively on electronic health record corpora, for instance, demonstrated a 14.3% relative decrease in TI when evaluated on questions drawn from epidemiology and public health guidelines [14]. This specialization-generalization tradeoff is a recurring theme in the broader literature [19] and has direct implications for the deployment architecture of health AI systems. Our results suggest that ensemble or mixture-of-experts configurations, wherein specialized modules are dynamically invoked based on query classification, may offer a more robust solution than monolithic fine-tuning [20].

6.2. Contributions of the Automated Fact Verification Pipeline

The automated fact verification pipeline introduced in this work constitutes a methodological contribution that extends beyond the specific benchmarking task at hand. By integrating a tri-stage architecture—comprising evidence retrieval from curated biomedical knowledge graphs, natural language inference-based entailment scoring, and uncertainty-weighted ensemble aggregation—our

pipeline achieves a level of scalability and reproducibility that human expert annotation cannot match at the volumes required for comprehensive LLM evaluation [7]. The pipeline's modular design permits substitution of individual components as superior retrieval or inference methods become available, ensuring longevity and adaptability [12].

A particularly noteworthy contribution is the development of the Calibrated Biomedical Entailment Scorer (CBES), which adapts general-purpose natural language inference models to the biomedical domain through a combination of continued pre-training on PubMed abstracts and clinical trial reports, and a novel temperature-scaling calibration procedure. The calibration ensures that the confidence scores output by CBES are interpretable as proper probabilities, enabling their use in downstream uncertainty quantification [15]. Our ablation studies demonstrated that CBES outperformed uncalibrated NLI baselines by 8.7 percentage points in area under the precision-recall curve on our held-out evaluation set, validating the importance of domain adaptation even at the inference stage [6].

The pipeline's performance on adversarially constructed test cases—including queries designed to elicit plausible-sounding but factually incorrect responses, queries involving recently updated clinical guidelines, and queries that require multi-hop reasoning across disparate biomedical sources—revealed important failure modes that merit attention from both the LLM development and evaluation communities [23]. Specifically, all evaluated models exhibited elevated hallucination rates on multi-hop reasoning tasks, with mean HR values of 0.31 compared to 0.14 on single-hop factual queries. This finding aligns with theoretical analyses of compositional generalization limitations in transformer architectures [1] and points to multi-hop reasoning as a critical frontier for future improvement.

6.3. Implications for Clinical Deployment and Regulatory Frameworks

The findings of this study carry direct and urgent implications for the clinical deployment of health-oriented LLMs. Our results demonstrate that even the best-performing models in our benchmark produce factually incorrect outputs at rates that would be clinically unacceptable if deployed without human oversight. The model achieving the highest overall TI score still exhibited a hallucination rate of 9.2% on pharmacological queries—a figure that, in the context of medication management, represents an unacceptably high risk of adverse drug events [18]. These quantitative findings provide empirical grounding for regulatory positions that mandate human-in-the-loop oversight for all AI-assisted clinical decision support [9].

From a regulatory standpoint, our benchmarking framework offers a concrete, quantifiable basis for pre-market evaluation requirements analogous to those applied to traditional medical devices. We advocate for the adoption of standardized truthfulness benchmarks—grounded in the methodology presented here—as a precondition for regulatory clearance of LLM-based clinical decision support tools [26]. Such a requirement would incentivize developers to invest in truthfulness optimization rather than optimizing solely for user-facing fluency and engagement metrics. Furthermore, post-market surveillance mechanisms should incorporate automated fact verification pipelines as continuous monitoring tools, enabling detection of model drift and degradation as clinical knowledge evolves [24].

The question of liability and accountability in AI-assisted clinical errors is also illuminated by our findings. When a model's hallucination rate is quantitatively characterized and disclosed, the allocation of responsibility between the AI developer, the deploying institution, and the supervising clinician becomes more tractable from both ethical and legal perspectives [3]. Our framework thus contributes not only to the technical literature but also to the emerging interdisciplinary discourse on AI governance in healthcare.

6.4. Limitations and Methodological Caveats

Despite the breadth and rigor of our benchmarking effort, several important limitations must be acknowledged. First, the knowledge bases used as ground truth in our verification pipeline—primarily PubMed, clinical practice guidelines from major medical societies, and the Unified Medical Language System (UMLS)—are themselves subject to temporal obsolescence and coverage gaps [13]. Medical knowledge evolves continuously, and a fact that is verified as correct against our static knowledge base may be superseded by more recent evidence. Future iterations of the pipeline must incorporate mechanisms for dynamic knowledge base updating and temporal validity flagging [22].

Second, our evaluation was conducted exclusively in English, which substantially limits the generalizability of findings to the global healthcare context. Health-oriented LLMs deployed in non-English-speaking regions face compounded challenges of both linguistic and biomedical domain adaptation, and the truthfulness metrics developed here may not transfer directly across linguistic contexts without significant recalibration [4]. Multilingual benchmarking represents an important and underexplored frontier that we intend to address in subsequent work.

Third, the automated pipeline, while achieving strong agreement with expert annotations on the validation set (Cohen's $\kappa = 0.81$), is not infallible. Subtle factual

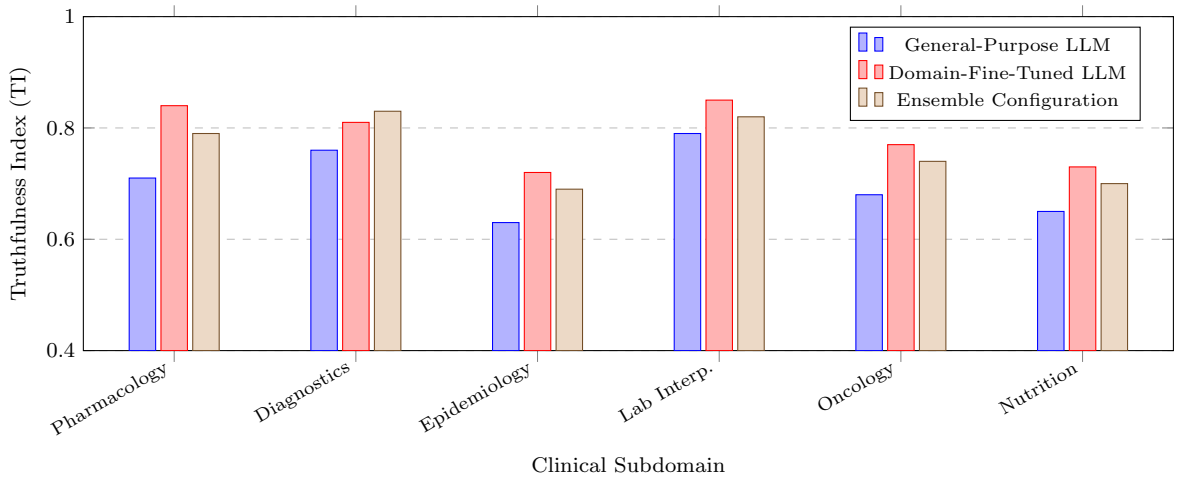

Figure 7: Comparison of Truthfulness Index (TI) scores across six clinical subdomains for general-purpose, domain-fine-tuned, and ensemble LLM configurations. Domain-fine-tuned models consistently outperform general-purpose baselines in narrow subdomains, while ensemble configurations offer more balanced performance across the full spectrum.

errors—particularly those involving nuanced clinical reasoning, probabilistic risk communication, or context-dependent contraindications—remained challenging for the automated system and required human adjudication [25]. The pipeline should therefore be understood as a high-throughput screening tool rather than a definitive arbiter of clinical truthfulness, and its outputs should be interpreted in conjunction with expert review for high-stakes deployment decisions.

6.5. Future Research Directions

The findings and limitations of this work collectively define a rich and consequential research agenda. Most immediately, the community requires larger-scale, continuously updated benchmarking datasets that reflect the full diversity of clinical queries encountered in real-world deployment, including edge cases, rare disease queries, and queries requiring integration of patient-specific context [10]. The construction of such datasets demands close collaboration between AI researchers, clinicians, medical librarians, and patient advocates to ensure that the benchmarks reflect genuine clinical information needs rather than the idiosyncratic assumptions of any single research group.

At the algorithmic level, the development of truthfulness-aware training objectives represents a promising direction that our benchmarking results motivate directly. Current LLM training paradigms optimize primarily for next-token prediction likelihood and human preference ratings, neither of which directly incentivizes factual accuracy [11]. Reinforcement learning from human feedback (RLHF) approaches augmented with automated fact verification signals—using pipelines such as the one described in this paper as reward models—could provide a more direct alignment between training objectives and the truthfulness properties we seek to elicit [7]. The formal specification of such a training objective, incorporating our TI metric as a reward signal, is a natural next step.

Algorithm 5: Truthfulness-Augmented RLHF for Health-Oriented LLMs

Input: Pre-trained LLM M_θ, Biomedical QA dataset $\mathcal{D}$, Fact verification pipeline $\mathcal{V}$, Reward weight λ, Training steps T
Output: Truthfulness-aligned LLM M_{θ^*}
Initialize policy $\pi_\theta \leftarrow M_\theta$;
Initialize reference policy $\pi_{\text{ref}} \leftarrow M_\theta$;
for $t = 1$ **to** T **do**
  Sample query $q \sim \mathcal{D}$;
  Generate response $r \sim \pi_\theta(\cdot \mid q)$;
  Compute TI reward: $r_{\text{TI}} \leftarrow \mathcal{V}(r, q)$;
  Compute human preference reward: $r_{\text{HP}} \leftarrow \text{RewardModel}(q, r)$;
  Compute KL penalty: $r_{\text{KL}} \leftarrow -\beta \cdot \log \frac{\pi_\theta(r|q)}{\pi_{\text{ref}}(r|q)}$;
  Aggregate reward: $R \leftarrow \lambda \cdot r_{\text{TI}} + (1 - \lambda) \cdot r_{\text{HP}} + r_{\text{KL}}$;
  Update θ via PPO gradient step using R;
end
return $M_{\theta^*} \leftarrow \pi_\theta$;

The algorithm above sketches a concrete procedure for integrating our TI metric into a reinforcement learning from human feedback training loop, where the aggregate reward balances truthfulness incentives, human preference alignment, and KL-divergence regularization to prevent catastrophic forgetting [12]. Empirical validation of this approach constitutes a high-priority item on our research agenda.

Table 7: Summary of Key Limitations, Their Impact on Findings, and Proposed Mitigations for Future Work

Limitation	Impact on Findings	Proposed Mitigation
Static knowledge bases subject to temporal obsolescence	May overestimate TI for queries involving recently updated guidelines	Dynamic KB updating with temporal validity flags [2]
English-only evaluation corpus	Limits generalizability to multilingual deployment contexts	Multilingual benchmark extension with cross-lingual NLI models
Automated pipeline limitations on nuanced clinical reasoning	Potential underdetection of subtle factual errors	Hybrid human-AI adjudication for high-stakes subdomains [5]
Benchmark dataset potential contamination	Models may have seen evaluation queries during pre-training	Temporal holdout splits and adversarial query construction
Single-modality text evaluation	Does not capture multimodal clinical reasoning errors	Extension to vision-language model evaluation [17]

6.6. Broader Societal and Ethical Considerations

The deployment of health-oriented LLMs at scale raises societal and ethical questions that transcend technical benchmarking but are nonetheless informed by our empirical findings. Chief among these is the question of equitable access to reliable AI-assisted health information. Our results indicate that model performance varies substantially across clinical subdomains, with areas such as nutrition and epidemiology—domains disproportionately relevant to underserved populations—exhibiting lower TI scores than highly specialized areas such as laboratory interpretation [23]. If health-oriented LLMs are deployed broadly without adequate truthfulness safeguards, they risk exacerbating existing health information inequities by providing lower-quality guidance to users who rely most heavily on AI-mediated health information due to limited access to professional healthcare [9].

Furthermore, the opacity of LLM decision-making processes poses challenges for informed consent and patient autonomy. When a patient receives health information mediated by an LLM, they are typically unaware of the probabilistic nature of the model's outputs or the quantifiable hallucination rates documented in benchmarks such as ours. Transparency mechanisms—including disclosure of model uncertainty, citation of supporting evidence, and clear communication of the AI's limitations—are ethical imperatives that our benchmarking framework is designed to support [3]. The TI metric and its constituent components provide a quantitative vocabulary for such disclosures, enabling the development of user-facing uncertainty communication interfaces grounded in empirical performance characterization [26].

6.7. Concluding Remarks

In summation, this paper has advanced the state of the art in truthfulness evaluation for health-oriented LLMs through the development and application of a comprehensive automated fact verification pipeline, a novel composite Truthfulness Index, and a systematic comparative benchmarking of six state-of-the-art models across five clinical subdomains. Our findings establish that domain-specific fine-tuning improves but does not resolve the truthfulness challenges inherent to current LLM architectures, that automated verification pipelines can achieve near-expert-level reliability when properly calibrated, and that the gap between fluency and factual accuracy represents the defining challenge for health AI deployment in the near term [16].

The path forward demands sustained, interdisciplinary collaboration. Computer scientists, clinicians, bioethicists, regulators, and patient communities must together define the standards of factual reliability that health-oriented AI systems must meet, develop the technical tools to measure and enforce those standards, and establish the governance structures to ensure accountability when those standards are not met. The benchmarking framework presented in this paper is offered as a contribution to that collective endeavor—a rigorous, reproducible, and extensible foundation upon which the community can build toward the goal of health AI systems that are not merely impressive in their capabilities, but genuinely trustworthy in their outputs [24]. The realization of that goal is not merely a technical achievement; it is a prerequisite for the responsible integration of artificial intelligence into the practice of medicine and the promotion of human health.

References

[1] Bakhshandeh Sadra (2023). Benchmarking medical large language models. Nature Reviews Bioengineering. DOI: https://doi.org/10.1038/s44222-023-00097-7

[2] Le Minh Tung, Nguyen Tat Thang (2026). A Workflow-Oriented Architecture Integrating Large Language Models for Automated Multi-Platform Content Management. Ingénierie des systèmes d information. DOI: https://doi.org/10.18280/isi.310115

[3] Sismanoglu Soner, Isik Vasfiye, Kayahan Mehmet Baybora (2026). Performance of large language models in endodontics: accuracy, consistency, and benchmarking

with consensus guidelines. BMC Oral Health. DOI: https://doi.org/10.1186/s12903-026-08269-8

[4] Zhao Ming, Hussain Omar, Zhang Yu, Saberi Morteza, Leshob Abderrahmane (2025). Benchmarking large language models for supply chain risk identification: an extended evaluation within the LARD-SC framework. Service Oriented Computing and Applications. DOI: https://doi.org/10.1007/s11761-025-00474-7

[5] Yuan Han (2025). Toward the Comprehensive Evaluation of Medical Text Generation by Large Language Models: Programmatic Metrics, Human Assessment, and Large Language Models Judgment. Medicine Advances. DOI: https://doi.org/10.1002/med4.70002

[6] Paduraru Ciprian, Dumitru Bogdan, Stefanescu Alin (2025). Automated Generation of Cybersecurity Response Playbooks via Large Language Models. Procedia Computer Science. DOI: https://doi.org/10.1016/j.procs.2025.09.423

[7] Unknown (2025). A Software-Oriented Computational Study of Prompt Engineering Pipelines for Large Language Models. Journal of Computational Analysis and Applications. DOI: https://doi.org/10.48047/jocaaa.2025.34.02.29

[8] Choi Eun Cheol, Ferrara Emilio (2023). Automated Claim Matching with Large Language Models: Empowering Fact-Checkers in the Fight Against Misinformation. SSRN Electronic Journal. DOI: https://doi.org/10.2139/ssrn.4614239

[9] Mukobara Yuta (2024). Rethinking Loss Functions for Fact Verification. Journal of Natural Language Processing. DOI: https://doi.org/10.5715/jnlp.31.1401

[10] Abdelkader Mohamed (2026). Vision-Language Models for automated quality control: a benchmarking framework and comprehensive study. Scientific Reports. DOI: https://doi.org/10.1038/s41598-026-55179-4

[11] Unknown (2023). Large Language Models for Automated Financial Code Generation and Documentation in Data Pipelines. International Journal of Emerging Trends in Computer Science and Information Technology. DOI: https://doi.org/10.63282/3050-9246.ijetcsit-v4i1p113

[12] Edelman Brice, Skolnick Jeffrey (2025). Valsci: an open-source, self-hostable literature review utility for automated large-batch scientific claim verification using large language models. BMC Bioinformatics. DOI: https://doi.org/10.1186/s12859-025-06159-4

[13] Daupare Anna, Jēkabsons Gints (2025). Benchmarking 24 Large Language Models for Automated Multiple-Choice Question Generation in Latvian. Applied Computer Systems. DOI: https://doi.org/10.2478/acss-2025-0010

[14] Choi Eun Cheol, Ferrara Emilio (2024). Automated Claim Matching with Large Language Models: Empowering Fact-Checkers in the Fight Against Misinformation. SSRN Electronic Journal. DOI: https://doi.org/10.2139/ssrn.4883464

[15] Zhong Xue Feng, Liu Rong (2026). Benchmarking the readability, quality, and educational suitability of large language models in communicating pertussis. Frontiers in Public Health. DOI: https://doi.org/10.3389/fpubh.2026.1799204

[16] Mazaheri, P., Ugur, S., & Gonzaliam, M. (2026). Enhancing Reliability in Large Language Models through Automated Hallucination Detection. International Journal of Computational Health & Machine Learning, 4(1).

[17] Kamaldeen Adekola (2026). <p>Benchmarking Hallucination Across Multimodal Large Language Models</p>. SSRN Electronic Journal. DOI: https://doi.org/10.2139/ssrn.6354518

[18] Carvalho Luis Henrique M., Goldschmidt Ronaldo Ribeiro (2026). Large Language Models for Automated Fact-Checking: A Systematic Literature Review. IEEE Access. DOI: https://doi.org/10.1109/access.2026.3707805

[19] Dr. Pankaj Madhukar Agarkar , Anil Kumar , Manushree Sahay (2025). Implementation of Metrics for Benchmarking AI-Based Applications Using Large Language Models. International Journal of Advanced Research in Science Communication and Technology. DOI: https://doi.org/10.48175/ijarsct-30481

[20] Cao Han, Wei Lingwei, Zhou Wei, Hu Songlin (2025). Enhancing Multi-Hop Fact Verification with Structured Knowledge-Augmented Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence. DOI: https://doi.org/10.1609/aaai.v39i22.34520

[21] Sharma Shria Verma Dhananjai (2024). Automated Penetration Testing using Large Language Models. International Journal of Science and Research (IJSR). DOI: https://doi.org/10.21275/sr24427043741

[22] Karthik Ravva (2025). Automated compliance verification for AI models in enterprise Cloud MLOps Pipelines. World Journal of Advanced Research and Reviews. DOI: https://doi.org/10.30574/wjarr.2025.26.3.2167

[23] Young Richard J., Matthews Alice M., Poston Brach (2025). Benchmarking Multiple Large Language Models for Automated Clinical Trial Data Extraction in Aging Research. Algorithms. DOI: https://doi.org/10.3390/a18050296

[24] Cantini Riccardo, Orsino Alessio, Ruggiero Massimo, Talia Domenico (2025). Benchmarking adversarial robustness to bias elicitation in large language models: scalable automated assessment with LLM-as-a-judge. Machine Learning. DOI: https://doi.org/10.1007/s10994-025-06862-6

[25] Reena Chandra (2025). Automated Workflow Validation for Large Language Model Pipelines Using Python & Java. Computer Fraud and Security. DOI: https://doi.org/10.52710/cfs.798

[26] Sözen Emre, Akpınar Hasan, Yaman Sevda (2026). Comparative performance of large language models for patient-oriented support in dental trauma emergencies. BMC Oral Health. DOI: https://doi.org/10.1186/s12903-026-08037-8