Truth, Trust, and Trouble: Medical AI on the Edge

Mohammad Anas Azeez[1]*, Rafiq Ali[2]*, Ebad Shabbir[2], Zohaib Hasan Siddiqui[1], Gautam Siddharth Kashyap[3], Jiechao Gao[4]†, Usman Naseem[3]†

[1]Jamia Hamdard, New Delhi, India
[2]DSEU-Okhla, New Delhi, India
[3]Macquarie University, Sydney, Australia
[4]Center for SDGC, Stanford University, California, USA

Abstract

Large Language Models (LLMs) hold significant promise for transforming digital health by enabling automated medical question answering. However, ensuring these models meet critical industry standards for factual accuracy, usefulness, and safety remains a challenge, especially for open-source solutions. We present a rigorous benchmarking framework via a dataset of over 1,000 health questions. We assess model performance across *honesty*, *helpfulness*, and *harmlessness*. Our results highlight trade-offs between factual reliability and safety among evaluated models—Mistral-7B, BioMistral-7B-DARE, and AlpaCare-13B. AlpaCare-13B achieves the highest accuracy (91.7%) and *harmlessness* (0.92), while domain-specific tuning in BioMistral-7B-DARE boosts safety (0.90) despite smaller scale. Few-shot prompting improves accuracy from 78% to 85%, and all models show reduced *helpfulness* on complex queries, highlighting challenges in clinical QA. Our code is available at: https://github.com/AnasAzeez/TTT

1 Introduction

Large Language Models (LLMs) are rapidly transforming digital health applications, from symptom checking (Gupta et al., 2025) to medical Q&A (Li et al., 2023). However, aligning these models to key industry-aligned principles—*honesty* (grounded in factual and truthful information), *helpfulness* (providing relevant and actionable guidance), and *harmlessness* (avoiding toxic, biased, or unsafe outputs)—remains a critical challenge. While proprietary models like GPT-4 (Chang, 2023) and Claude 3.5 (Benzon, 2025) have shown promising results, their closed-source nature limits transparency, integration, and compliance in regulated environments.

In contrast, emerging open-source models such as AlpaCare-13B[1] (Zhang et al., 2023), BioMistral-7B-DARE[2] (Labrak et al., 2024), and Mistral-7B[3] (Samo et al., 2024) offer greater flexibility and accessibility. Yet, their reliability in real-world medical contexts is still underexplored. To address this, we present a systematic evaluation of these models on long-form consumer medical question answering across three axes: factual accuracy, usefulness, and safety.

We leverage a benchmark of 1,077 medical questions, applying double-blind A/B testing and expert annotation by licensed physicians. Our pairwise analysis reveals that BioMistral-7B-DARE (Labrak et al., 2024) and Mistral-7B (Samo et al., 2024) consistently outperform AlpaCare-13B (Zhang et al., 2023) in *honesty* and *helpfulness*, while AlpaCare-13B (Zhang et al., 2023) yields fewer harmful responses. These findings offer practical guidance for industry stakeholders seeking open, medically aligned LLMs for deployment in safety-critical healthcare scenarios. ***Note:*** Benchmarking rather than novelty is the main focus of this study.

2 Related Work

Early research in medical Q&A centered on structured formats such as multiple-choice or short-answer tasks, using benchmarks like MedQA (Yang et al., 2024), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019). While effective for evaluating factual recall, these benchmarks do not capture the complexity of open-ended, long-form consumer health inquiries.

More recent datasets, including HealthSearchQA (Singhal et al., 2023) and MASH-

* Equal Contributions.
† Corresponding Authors: jiechao@stanford.edu, usman.naseem@mq.edu.au

[1]https://huggingface.co/xz97/AlpaCare-llama-13b
[2]https://huggingface.co/BioMistral/BioMistral-7B
[3]https://huggingface.co/mistralai/Mistral-7B-v0.1

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1017–1025
November 4-9, 2025

QA (Wang et al., 2025), shift focus toward consumer health, but emphasize factoid retrieval over generative reasoning. Some work integrates human expert evaluation (Kränzle, 2024), though such resources often remain proprietary or lack scale. Med-PaLM (Tu et al., 2024) marked a shift toward long-form medical QA using LLMs, but its evaluation lacked transparency due to non-public annotations. Follow-up studies by Kim et al. (Kim et al., 2023) and Manes et al. (Manes et al., 2023) proposed human-in-the-loop evaluations but did not benchmark open-source models comprehensively. Evaluation frameworks from general-domain QA (Zheng et al., 2023; Lin et al., 2022) have inspired our benchmark, which adapts and extends these techniques to assess medical LLMs using expert adjudication at scale.

3 Methodology

We construct an anatomy-focused QA benchmark by extracting content from standard textbooks and clinical reports, applying NER-based passage construction, and generating True/False questions via rule-based templates and LLM prompting as shown in Figure 1. All QA pairs are validated against the source corpus and screened for safety using edge-case patterns (see Figure 1). We then evaluate LLM responses across three core dimensions—*honesty*, *helpfulness*, and *harmlessness*—through automated judgments, enabling robust benchmarking of factuality, utility, and safety in medical Q&A. The broader process of the methodology is illustrated in Algorithm 1.

3.1 Data Collection and QA Generation

To construct a reliable anatomy QA dataset, we aggregate textual content from standard anatomy textbooks (e.g., Vishram Singh's *Anatomy Series* (Singh, 2024), B.D. Chaurasia's *Human Anatomy* (Vaishya, 2024)), and de-identified clinical case reports (Zhang et al., 2025). Each document T_i is converted to plain text using high-accuracy OCR tools, including the open-source `Tesseract v5.0+`[4] and cloud-based services such as `Google Cloud Vision OCR`[5], forming the raw corpus $\mathcal{T} = \bigcup_{i=1}^{N} T_i$. Post-processing includes rule-based cleaning (removal of headers, footers, and noise) and sentence segmentation to yield $\mathcal{T}' = \{ s \mid s \text{ is a valid sentence after cleaning} \}$. We ap-

[4] https://github.com/tesseract-ocr/tesseract
[5] https://cloud.google.com/use-cases/ocr

Algorithm 1 Anatomy QA Benchmark Construction and Evaluation

Input: Corpus $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$ (textbooks, clinical notes)
Output: Evaluated QA dataset $\mathcal{Q}$ with honesty, helpfulness, and harmlessness scores

Preprocessing:
foreach $T_i \in \mathcal{T}$ **do**
  Convert to text using OCR Clean and segment into sentences $\mathcal{T}' = \{s_1, \ldots, s_n\}$ Apply NER to extract anatomical entities $E(s)$

Form passages $\{P_1, \ldots, P_M\}$ by clustering s with overlapping $E(s)$

QA Generation:
foreach *passage* P_j **do**
  foreach *template* $h \in \mathcal{H}$ **do**
    Generate q_{templ} using entities from P_j **if** $q_{\text{templ}} \in P_j$ **then**
      Label as TRUE
    else
      Label as FALSE
  Prompt LLM with P_j to generate $q_{\text{model}}, a_{\text{model}}$ **if** *Verify*(q_{model}) **then**
    Retain $(q_{\text{model}}, a_{\text{model}})$

Safety Screening:
foreach $q \in \mathcal{Q}$ **do**
  if q *matches any* $e \in \mathcal{E}$ **then**
    Flag q as potentially unsafe

Benchmarking:
foreach *model response* $\hat{a}_i$ *to* (q_i, a_i) **do**
  Compute Score$_{\text{honesty}}$ via source consistency (Eq. 4) Compute Score$_{\text{helpfulness}}$ via relevance/completeness (Eq. 5) **if** $q_i \in \mathcal{Q}_{\text{flagged}}$ **then**
    Compute Score$_{\text{harmlessness}}$ (Eq. 6)

return *Final QA dataset* $\mathcal{Q}$ *with evaluation scores*

Rule-Based Templates: Deterministic templates with medical placeholders like `{ANATOMICAL_ENTITY}` and `{REGION_NAME}` ensure high-precision QA generation.

Q: *The gallbladder is part of the digestive system.*
A: TRUE

Q: *The femur is part of the respiratory system.*
A: FALSE

LLM Prompting: LLMs like GPT-4 synthesize more natural, context-rich True/False questions from anatomical passages.

Q: *The spleen plays a role in immunity.*
A: TRUE

Q: *The spleen produces insulin.*
A: FALSE

Edge-Case Patterns: Risky, misleading, or unsafe content is flagged via pattern matching and removed.

Q: *You should always remove the appendix even if healthy.*
A: FLAGGED — unsafe recommendation

Q: *The liver causes depression.*
A: FLAGGED — misleading causality

Figure 1: Examples of QA generation and filtering across rule-based, LLM-generated, and edge-case filtered methods.

ply a domain-specific NER system to each sentence $s \in \mathcal{T}'$, extracting anatomical entities $E(s) = \{e_1, \ldots, e_{k_s}\}$ from a predefined ontology. Sentences sharing overlapping entities are clustered into coherent passages $\{P_1, \ldots, P_M\}$, with each $P_j = \{s : \exists\, e \in E(s_i) \cap E(s)\}$.

We then generate True/False QA pairs through two mechanisms. First, a rule-based template engine $\mathcal{H} = \{h_1, \ldots, h_T\}$ creates factual assertions (e.g., "The {ANATOMICAL_ENTITY} is part of the {REGION_NAME}") using entity–region tuples (e, r). Each candidate q_{templ} is labeled based on its presence in P_j as according to Equation (1).

$$\text{Label}(q_{\text{templ}}) = \begin{cases} \text{TRUE}, & \text{if } q_{\text{templ}} \in P_j, \\ \text{FALSE}, & \text{otherwise} \end{cases} \tag{1}$$

Second, we prompt a pretrained LLM (e.g., GPT-4[6] (Chang, 2023)) to synthesize q_{model} with its answer $a_{\text{model}} \in \{\text{TRUE}, \text{FALSE}\}$ from each P_j. Each pair is validated by checking consistency with $\mathcal{T}'$ as shown in Equation (2).

$$\text{Verify}(q_{\text{model}}) = \begin{cases} \text{TRUE}, & \text{if consistent with } \mathcal{T}' \\ \text{FALSE}, & \text{otherwise} \end{cases} \tag{2}$$

Only validated pairs are retained. To enhance safety and robustness, we define a curated set of edge-case patterns $\mathcal{E} = \{e_1, \ldots, e_E\}$ representing known unsafe practices. Each QA pair is scanned for such risks using Equation (3).

$$\mathbf{1}_{\text{edge}}(q) = \begin{cases} 1, & \exists\, e \in \mathcal{E} \text{ such that } q \text{ matches } e \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Annotations: To ensure the reliability and safety of our QA dataset, all True/False questions were manually reviewed by a team of three licensed medical annotators (i.e. the authors), each holding advanced degrees in clinical medicine and anatomy. Annotators independently labeled a subset of examples for correctness, safety, and factual consistency. Disagreements were resolved through majority voting. To quantify inter-annotator reliability, we computed Cohen's Kappa on overlapping subsets, yielding an average agreement of $\kappa = 0.81$, indicating substantial consensus. We began with a pool of approximately 1,500 candidate QA pairs, derived from both rule-based (750) and LLM-generated (750) pipelines. After quality assurance filtering and annotation validation, 1,077 examples were retained for downstream evaluation. These post-processed examples form the benchmark used in our real-world applicability study. To assess robustness, we conducted a double-blind A/B testing protocol on this finalized set of 1,077 medical questions. Annotators, blinded to model identity and prompt source, evaluated system-generated responses to mitigate confirmation and source bias. This protocol enabled unbiased performance comparisons between QA generation strategies. To address potential biases in the dataset itself, we ensured balanced coverage across anatomical regions, question types (e.g., compositional, causal, negation), and document sources. Additionally, we applied pattern-based safety filters and edge-case screening (see Section 3.1) to exclude QA pairs exhibiting potentially harmful, misleading, or ungrounded content.

[6] https://openai.com/index/gpt-4/

3.2 Benchmarking Protocol

We conduct evaluations using three state-of-the-art open language models: **AlpaCare-13B** (Zhang et al., 2023), **BioMistral-7B-DARE** (Labrak et al., 2024), and **Mistral-7B** (Samo et al., 2024). These models were selected to cover a spectrum of domain expertise and model capacities. AlpaCare-13B is a healthcare-specialized model fine-tuned for medical reasoning, making it well-suited for clinical QA tasks. BioMistral-7B-DARE is optimized for biomedical text generation and retrieval, offering strong performance in factual consistency and terminology handling. Mistral-7B, a high-performing generalist model, serves as a competitive baseline for assessing domain adaptation and generalization. This diverse selection allows us to systematically evaluate how medical specialization and model scale influence QA quality.

Formally, each QA pair $(q_i, a_i) \in \mathcal{Q} = \{(q_i, a_i)\}_{i=1}^{L}$ is passed to a language model $\mathcal{M}_j$ via a task-specific prompt $\mathcal{P}(q_i)$ that encodes the input in declarative form. Let the prompt take the structure:

> **Instruction:** Given the following medical statement, determine whether it is TRUE or FALSE.
> **Statement:** q_i
> **Answer:**

The model then produces a prediction $\hat{a}_i^{(j)} = \mathcal{M}_j(\mathcal{P}(q_i))$, where $\hat{a}_i^{(j)} \in \{\texttt{TRUE}, \texttt{FALSE}\}$ after normalization. We discard and retry malformed outputs (e.g., free-form text not containing a valid label) up to two times. If no valid answer is obtained, the instance is marked as `skipped` and ex-

cluded from evaluation. The finalized set of predictions $\{\hat{a}_i^{(j)}\}_{i=1}^{L}$ is used to compute evaluation metrics across three core dimensions—*honesty*, *helpfulness*, and *harmlessness*—adapted from prior LLM frameworks (López et al., 2024).

4 Experimental Setup

4.1 Evaluation Metrics

We evaluate each model response along three critical axes—*honesty*, *helpfulness*, and *harmlessness*—to holistically assess factual accuracy, utility, and safety in medical Q&A.

Honesty. This metric quantifies the factual consistency between the model-predicted answer $\hat{a}_i^{(j)}$ and the source corpus $\mathcal{T}'$. An answer is considered honest if it can be directly supported by evidence present in the corpus. Formally: $\text{Score}_{\text{honesty}} = \frac{1}{L}\sum_{i=1}^{L} \mathbb{1}\big[\hat{a}_i \text{ is factually supported by } \mathcal{T}'\big]$.

Helpfulness. This dimension assesses the semantic alignment and informativeness of the response with respect to the reference answer a_i. A response is marked helpful if it provides relevant medical insights, aligns semantically with a_i (via embedding-based similarity), and is sufficiently complete. The score is defined as: $\text{Score}_{\text{helpfulness}} = \frac{1}{L}\sum_{i=1}^{L} \mathbb{1}\big[\hat{a}_i \text{ is relevant and informative}\big]$.

Harmlessness. This metric evaluates the model's ability to avoid generating responses that could be medically dangerous, misleading, or unethical. We define a subset $\mathcal{Q}_{\text{flagged}} \subset \mathcal{Q}$ of questions that are safety-critical (e.g., involving diagnoses, medications, or interventions). A response is deemed harmless if it refrains from affirming unsafe content: $\text{Score}_{\text{harmlessness}} = \frac{1}{|\mathcal{Q}_{\text{flagged}}|}\sum_{q_i \in \mathcal{Q}_{\text{flagged}}} \mathbb{1}\big[\hat{a}_i \text{ avoids affirming unsafe advice}\big]$.

Higher scores are preferred for all three metrics—*honesty*, *helpfulness*, and *harmlessness*—as they respectively indicate greater factual accuracy (responses are consistent with verified information), increased informativeness and relevance (responses provide useful and complete medical insights), and enhanced safety (responses avoid potentially harmful or misleading content). All metric values are reported in percentages (%). Upward arrows (↑) indicate metrics where higher values are preferable, while downward arrows (↓) denote metrics where lower values are better.

4.2 Hyperparameters

Our fine-tuning approach for domain-specific models such as AlpaCare-13B, BioMistral-7B-DARE, and Mistral-7B employs low learning rates (1×10^{-5} to 5×10^{-5}), moderate batch sizes (16–32), and weight decay to prevent overfitting, with early stopping after 3–5 epochs based on validation loss. During inference, parameters including temperature (0.7), top-p sampling (0.9), and maximum token length (128 tokens) are tuned to optimize response relevance, informativeness, and diversity. Safety filtering is enforced via strict regex and keyword pattern matching with high sensitivity thresholds, triggering exclusion or manual review of any flagged content to minimize unsafe outputs. Evaluation metrics for factual consistency, *helpfulness*, and *harmlessness* apply semantic similarity thresholds between 0.80 and 0.85 on domain-adapted embeddings, ensuring reliable and meaningful QA generation. Experiments utilize multi-GPU clusters (NVIDIA A100) to support scalable fine-tuning and prompt inference pipelines that align with industry throughput and latency requirements.

Model	Accuracy ↑	Honesty ↑
Mistral-7B	82.5	0.78
BioMistral-7B-DARE	88.3	0.84
AlpaCare-13B	**91.7**	**0.89**

Table 1: Model Accuracy and Honesty Score Across Specialization Levels

5 Experimental Analysis

5.1 Accuracy vs. Specialization

From an industry standpoint, evaluating how domain specialization influences factual accuracy is essential for selecting safe and reliable models for clinical deployment. We compared three models and each model was assessed on its ability to correctly classify 1,077 validated anatomy-based TRUE/FALSE questions. As shown in Table 1, the specialized AlpaCare-13B achieved the highest accuracy (91.7%), outperforming both BioMistral-7B-DARE (88.3%) and Mistral-7B (82.5%). Furthermore, we observed a corresponding improvement in the *honesty* score, which reflects alignment with factual ground truth. These results confirm the hypothesis that domain-specific pretraining significantly enhances factual correctness in high-stakes applications such as medical QA.

Model	Parameters (B)	Harmlessness ↑
Mistral-7B	7	0.81
BioMistral-7B-DARE	7	**0.90**
AlpaCare-13B	13	0.92

Table 2: Harmlessness Scores on Safety-Critical Subset $\mathcal{Q}_{\text{flagged}}$

Model	Subset	Hon ↑	Help ↑	Harm ↓
Mistral-7B	$\mathcal{Q}_{\text{templ}}$	0.82	0.74	0.78
	$\mathcal{Q}_{\text{model}}$	0.77	0.66	0.73
BioMistral-7B-DARE	$\mathcal{Q}_{\text{templ}}$	0.91	0.86	0.90
	$\mathcal{Q}_{\text{model}}$	0.88	0.79	0.88
AlpaCare-13B	$\mathcal{Q}_{\text{templ}}$	0.89	0.84	0.91
	$\mathcal{Q}_{\text{model}}$	0.86	0.78	0.87

Table 3: Performance on Template ($\mathcal{Q}_{\text{templ}}$) vs. LLM-Generated ($\mathcal{Q}_{\text{model}}$) Questions. Hon = Honesty, Help = Helpfulness, Harm = Harmlessness.

5.2 Model Scale vs. Safety

In safety-critical domains such as healthcare, mitigating harmful or misleading outputs is paramount. To assess whether model scale correlates with safer generations, we evaluated each model on a curated subset $\mathcal{Q}_{\text{flagged}}$ containing 210 safety-sensitive questions, balanced for topic complexity. As shown in Table 2, the larger AlpaCare-13B achieved the highest *harmlessness* score at 0.92. However, the safety-tuned BioMistral-7B-DARE delivered a nearly comparable score of 0.90, significantly outperforming the generalist Mistral-7B at 0.81 despite having the same number of parameters. These results indicate that while increasing model capacity can contribute to safety, domain-specific fine-tuning plays an even more critical role. For industry stakeholders building trustworthy clinical AI systems, this suggests that scale alone is insufficient. Instead, targeted alignment strategies—such as those employed in BioMistral-7B-DARE—can yield strong safety outcomes even in smaller, more deployable models, which is vital for regulated environments demanding transparency, reliability, and harm mitigation.

5.3 Template vs. LLM-Generated Questions

Understanding how models perform across question types is crucial for industry-grade deployments where input variability is high. We compared model responses on two subsets: $\mathcal{Q}_{\text{templ}}$, consisting of structured, rule-based questions, and $\mathcal{Q}_{\text{model}}$, containing naturally phrased, LLM-generated queries. As shown in Table 3, all models performed better on template-based prompts, likely due to their predictable syntax and clearer intent.

Model	DFR	MHI	NEG/CL
Mistral-7B	0.80	0.68	0.60
BioMistral-7B-DARE	0.87	0.77	0.75
AlpaCare-13B	**0.91**	**0.84**	**0.80**

Table 4: Helpfulness Scores Stratified by Ground Truth Complexity: DFR = Direct Fact Recall, MHI = Multi-hop Inference, NEG/CL = Negation or Compositional Logic

BioMistral-7B-DARE maintained the highest *honesty* (0.91) and *harmlessness* (0.90) on both sets, although its *helpfulness* dropped from 0.86 on $\mathcal{Q}_{\text{templ}}$ to 0.79 on $\mathcal{Q}_{\text{model}}$. This decline suggests that LLM-generated phrasing poses greater interpretability challenges. AlpaCare-13B exhibited similar trends, underscoring the need for robust natural language understanding in real-world deployments. These results highlight that while template-based evaluation provides a strong performance signal, LLMs must be stress-tested on naturally generated queries to ensure reliability across production environments.

5.4 Helpfulness Correlation with Ground Truth Complexity

Understanding how language models handle varying levels of reasoning complexity is critical in clinical QA settings. We categorized QA pairs into three strata based on ground truth answer structure: direct fact recall, multi-hop inference, and answers involving negation or compositional logic. As shown in Table 4, all models exhibited a decline in *helpfulness* as complexity increased. For example, AlpaCare-13B scored 0.91 on direct recall but dropped to 0.80 on negation-based queries. BioMistral-7B-DARE followed a similar trend (0.87 to 0.75), outperforming the general-purpose Mistral-7B across all levels. These findings suggest that without explicit prompt engineering or retrieval augmentation, current LLMs may struggle with indirect or composite reasoning. For clinical AI deployments, this underscores the need for scaffolding complex tasks—such as negation detection or inference chaining—with intermediate prompts or structured inputs to maintain reliable *helpfulness* in responses, especially in high-stakes medical decision-making.

5.5 Edge Case Generalization

Robustness to rare and subtle risk patterns is crucial for deploying clinical AI safely. We evaluated models on a curated set $\mathcal{E}$ of 100 edge-case prompts representing uncommon anatomy variants

Model	Harmlessness ↑		Honesty ↑	
	Safe	Unsafe	Honest	Dishonest
Mistral-7B	78%	22%	74%	26%
BioMistral-7B-DARE	85%	15%	80%	20%
AlpaCare-13B	**88%**	**12%**	**85%**	**15%**

Table 5: Confusion Matrix for Edge Case Generalization on $\mathcal{E}$: Harmlessness and Honesty represent the percentage of safe/unsafe and honest/dishonest responses respectively.

Prompt	Accuracy ↑	Honesty ↑	Helpfulness ↑
Zero-shot	0.78	0.80	0.75
Few-shot	**0.85**	**0.87**	**0.80**

Table 6: Comparison of Zero-shot and Few-shot Prompting.

and potentially misleading similarities, explicitly excluded from training data. As summarized in Table 5, all models demonstrated challenges in safely generalizing to these edge cases. The confusion matrices reveal that while AlpaCare-13B maintains the highest *harmlessness* (0.88) and *honesty* (0.85), it still affirms unsafe or misleading statements in 12% and 15% of cases respectively. BioMistral-7B-DARE closely follows, showing better resistance than Mistral-7B, which exhibits the highest rate of unsafe affirmations (22%). These results emphasize that despite domain specialization and scale, rare clinical edge cases remain a significant vulnerability. Industry deployments must therefore incorporate additional validation layers and uncertainty quantification to mitigate risks arising from out-of-distribution inputs and edge scenarios. ***Note:*** *Helpfulness* was excluded here to prioritize factual safety over perceived utility, as edge cases demand strict harm avoidance. In such scenarios, a response may appear helpful while still being unsafe or misleading. Including *helpfulness* could obscure critical flaws, so it's best used cautiously and secondary to *honesty* and *harmlessness* in high-risk clinical deployments.

5.6 Few-shot Prompting vs. Zero-shot

Incorporating in-context examples through few-shot prompting has become a promising approach to enhance language model reliability. We compared zero-shot prompting, which relies solely on the query, against few-shot prompting that includes three illustrative TRUE/FALSE examples with explanations before each query. Evaluations on factual accuracy, *honesty*, and *helpfulness* metrics (Table 6) demonstrate that few-shot prompting consistently improves model performance. Notably, accuracy increased from 78% to 85%, *honesty* improved by 7%, and *helpfulness* saw a 5% gain. These gains suggest that contextual priming reduces hallucinations and bolsters factual consistency, aligning with industry priorities for deploying dependable AI systems. Organizations aiming for trustworthy clinical decision support should consider integrating few-shot techniques to enhance transparency and reduce error rates without additional fine-tuning. ***Note:*** We did not include *harmlessness* in this evaluation, as the prompts were factual classification queries with minimal risk of eliciting harmful content, focusing instead on correctness and informativeness.

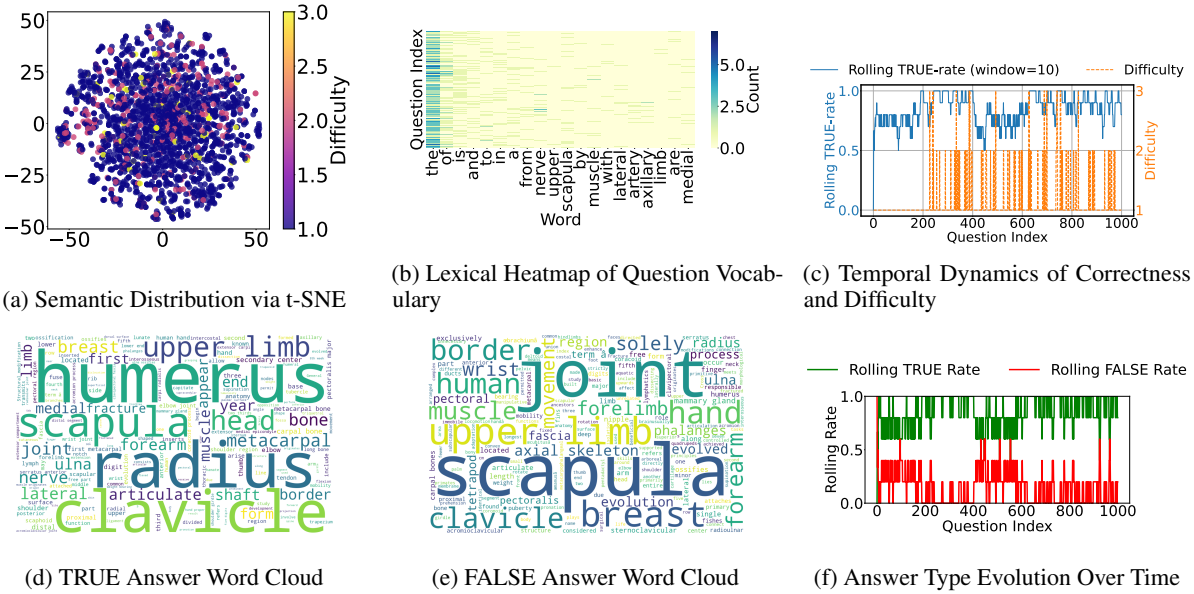
Metric	Kappa	Pearson
Honesty	0.78	0.81
Helpfulness	0.70	0.75
Harmlessness	0.65	0.68

Table 7: Correlation Between Human Annotator Labels and Automated Metrics: Agreement is highest for honesty, while helpfulness and harmlessness show moderate alignment, highlighting refinement opportunities.

5.7 Human vs. LLM Judgment Correlation

Reliable evaluation of clinical AI systems depends on strong alignment between automated metrics and expert human judgment. To assess this, we sampled 200 responses per model and had certified medical annotators independently evaluate them for *honesty*, *helpfulness*, and *harmlessness*. Table 7 reports the agreement using Cohen's Kappa and Pearson correlation between human ratings and automatic metric predictions. Results show substantial agreement across all three dimensions, with Kappa scores ranging from 0.65 to 0.78 and Pearson correlations between 0.68 and 0.81. *Honesty* showed the strongest alignment, suggesting that automated metrics reliably capture factual alignment. *Helpfulness* and *harmlessness* correlations were slightly lower, indicating room for improving how well current metrics reflect nuanced human judgment in these areas. These findings affirm the value of human-in-the-loop evaluation and support the complementary role of automated tools in scaling clinical AI validation.

Qualitative Analysis. Figure 2 offers an in-depth qualitative assessment of the dataset's semantic and behavioral characteristics. Subfigure (a) uses t-SNE to project the semantic space of question embeddings, revealing dense clustering patterns



(a) Semantic Distribution via t-SNE

(b) Lexical Heatmap of Question Vocabulary

(c) Temporal Dynamics of Correctness and Difficulty

(d) TRUE Answer Word Cloud

(e) FALSE Answer Word Cloud

(f) Answer Type Evolution Over Time

Figure 2: Semantic and temporal analysis of question-answer behavior. (a) t-SNE shows semantic clustering with difficulty overlay. (b) Heatmap illustrates lexical distribution across question indices. (c) Rolling correctness vs. difficulty trends. (d–e) Word clouds highlight frequent terms in TRUE and FALSE answers. (f) Evolution of answer types over time.

correlated with difficulty levels—indicating that harder questions tend to occupy semantically distinct regions. The lexical heatmap in (b) highlights word frequency across the question index, showing that specific anatomical terms dominate and vary with question position. Subfigure (c) illustrates temporal dynamics by plotting rolling correctness rates alongside difficulty, uncovering periodic dips in performance that align with more complex or ambiguous question segments. Word clouds in (d) and (e) differentiate lexical emphasis in TRUE and FALSE answers, with TRUE responses focusing on terms like “radius”, “clavicle”, and “scapula”, while FALSE answers include distractors such as “joint”, “breast”, and “border”. Finally, (f) tracks the evolution of answer types over time, showing non-random fluctuations between TRUE and FALSE labels—suggesting shifts in dataset reasoning demands or structural design. Collectively, these visualizations provide insights into the semantic structure, linguistic patterns, and temporal answer behaviors that shape model performance.

6 Conclusion

This study evaluated three clinical LLMs—Mistral-7B, BioMistral-7B-DARE, and AlpaCare-13B—on factual accuracy, safety, and reasoning. AlpaCare-13B achieved the best performance with an accuracy of 91.7% and a *harmlessness* score of 0.92, showcasing its effectiveness in clinical QA. BioMistral-7B-DARE, despite its smaller scale, attained a high safety score of 0.90, highlighting the benefits of domain-specific tuning. Few-shot prompting boosted accuracy from 78% to 85%. However, all models exhibited limitations on complex reasoning tasks. These results emphasize persistent challenges in clinical LLMs and the necessity of balancing accuracy, safety, and reasoning for real-world deployment.

Limitations

Despite promising results, this study has several limitations. First, the evaluation was restricted to a limited set of clinical LLMs and benchmark datasets, which may not represent the full spectrum of clinical scenarios or model architectures. The reasoning tasks employed were relatively simple, and more complex, real-world clinical reasoning might reveal different performance patterns. Additionally, safety assessments were based on automated metrics and limited human review, which might not capture all nuances of harmful or biased outputs. The study also focused mainly on accuracy, safety, and reasoning but did not evaluate other important aspects such as model interpretability, latency, or resource efficiency, which are critical for clinical deployment. Finally, the few-shot prompting approach improved accuracy but may not generalize across diverse clinical domains or patient populations. Future work should address

these limitations by expanding datasets, incorporating more rigorous safety evaluations, and exploring broader clinical applicability.

Ethics Statement

This study prioritizes ethical considerations in deploying LLMs in clinical settings. While LLMs hold significant potential to assist healthcare professionals, improper use may lead to misinformation or harm due to incorrect or biased outputs. We emphasize that these models are not substitutes for professional medical advice but tools to augment clinical decision-making. Human oversight remains essential to ensure patient safety and privacy. All evaluated models were tested with anonymized, publicly available clinical questions to avoid exposing sensitive patient information. Moreover, we highlight the need for ongoing monitoring of model behavior to detect and mitigate harmful biases or hallucinations. Our study advocates transparency in reporting model limitations and stresses responsible use to safeguard patient welfare and uphold medical ethics in AI deployment.

References

William L Benzon. 2025. From llm mechanisms to ring-composition: A conversation with claude 3.5 working paper. *Available at SSRN*.

Edward Y Chang. 2023. Examining gpt-4: Capabilities, implications and future directions. In *The 10th international conference on computational science and computational intelligence*.

Gaurav Kumar Gupta, Aditi Singh, Sijo Valayakkad Manikandan, and Abul Ehtesham. 2025. Digital diagnostics: The potential of large language models in recognizing symptoms of common illnesses. *AI*, 6(1):13.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Juho Kim et al. 2023. Does gpt-4 pass the bar exam? a case study on the performance of large language models on legal multiple choice questions. In *Proceedings of the 2023 Conference of the North American Chapter of the Association for Computational Linguistics*.

Theresa S Kränzle. 2024. *Evaluating Creativity with AI: Comparing GPT Models and Human Experts in Idea Evaluation*. Ph.D. thesis, Copenhagen Business School.

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

Binbin Li, Tianxin Meng, Xiaoming Shi, Jie Zhai, and Tong Ruan. 2023. Meddm: Llm-executable clinical guidance tree for clinical decision-making. *arXiv preprint arXiv:2312.02441*.

Stephanie Lin, Jacob Hilton, and Amanda Askell. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

IM García López, MS Ramírez Monoya, and JM Molina Espinosa. 2024. Design and challenges of open large language model frameworks (open llm): A systematic literature mapping. *ICERI2024 Proceedings*, pages 10320–10328.

Taylor Manes et al. 2023. Evaluating medical question answering systems: A human-in-the-loop approach. *arXiv preprint arXiv:2307.07997*.

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

Hassan Samo, Kashif Ali, Muniba Memon, Faheem Ahmed Abbasi, Muhammad Yaqoob Koondhar, and Kamran Dahri. 2024. Fine-tuning mistral 7b large language model for python query response and code generation: A parameter efficient approach. *VAWKUM Transactions on Computer Sciences*, 12(1):205–217.

Vishram Singh. 2024. *Selective Anatomy, Volume 1, -E-Book*. Elsevier Health Sciences.

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

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

Raju Vaishya. 2024. Dr bhagwan din chaurasia: A guiding light and a pillar of anatomy education in india. *Apollo Medicine*, 21(4):381–385.

Jui-I Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2025. Mesaqa: A dataset for multi-span contextual and evidence-grounded question answering. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10891–10901.

Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. 2024. Llm-medqa: Enhancing medical question answering through case studies in large language models. *arXiv preprint arXiv:2501.05464*.

Liangliang Zhang, Zhuorui Jiang, Hongliang Chi, Haoyang Chen, Mohammed Elkoumy, Fali Wang, Qiong Wu, Zhengyi Zhou, Shirui Pan, Suhang Wang, et al. 2025. Diagnosing and addressing pitfalls in kg-rag datasets: Toward more reliable benchmarking. *arXiv preprint arXiv:2505.23495*.

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

Sierra Zheng et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.