U-MIRAGE: Benchmarking Chain-of-Thought Reasoning for Urdu Medical QA

Ali Faheem[1,*], **Faizad Ullah**[1,*], **Muhammad Hammad**[1], **Ahmed Hassan**[1],
Muhammad Sohaib Ayub[2], **Asim Karim**[1],

[1] Department of Computer Science, Lahore University of Management Sciences, Lahore, Pakistan
{ali.faheem,20030057,hammad.muhammad,26100308,akarim}@lums.edu.pk
[2] Data Science Institute, University of Galway, Galway, Ireland
muhammadsohaib.ayub@universityofgalway.ie

Correspondence: ali.faheem@lums.edu.pk

Abstract

Medical AI systems increasingly rely on large language models (LLMs), yet their deployment in linguistically diverse regions remains unexplored. We address this gap by introducing U-MIRAGE, the first medical question-answering benchmark for Urdu and Roman Urdu. Urdu is the 11th most spoken language[1] (with over 246 million speakers) worldwide. Our systematic evaluation of six state-of-the-art LLMs reveals three main findings. (1) 6% to 10% drop in performance when moving from English to Urdu variants, even though medical knowledge should theoretically transfer across languages. (2) Chain-of-Thought (CoT) prompting improves small models by 8% to 20%, while surprisingly the larger models' performance degraded by up to 3%. (3) Quantized small models fail catastrophically in low-resource languages, achieving near-random accuracy regardless of various prompting strategies. These findings challenge core assumptions about multilingual medical AI systems. Roman Urdu consistently outperforms standard Urdu script, suggesting orthographic alignment with pre-training data matters more than linguistic proximity. CoT prompting effectiveness depends critically on model architecture rather than task complexity alone. CoT prompting effectiveness depends critically on model architecture rather than task complexity alone. Our contributions are threefold: (1) U-MIRAGE, (2) systematic benchmarking of LLMs for Urdu and Roman Urdu medical reasoning, and (3) empirical analysis of CoT prompting in low-resource contexts. Our code and datasets are publicly available[2].

1 Introduction

Large language models have transformed natural language processing, achieving human-level performance across diverse tasks from language understanding to complex reasoning (Brown et al., 2020; Chowdhery et al., 2023). However, this progress has been uneven across languages. While models perform well on English medical benchmarks (Singhal et al., 2025; Thirunavukarasu et al., 2023), their capabilities in other languages remain limited (Joshi et al., 2020; Ahuja et al., 2023). This gap creates hurdles for billions of people who communicate in languages with limited resources. The challenge is critical in healthcare systems, where accurate language understanding can be a matter of life and death. Therefore, it is essential to understand the reasons behind this gap.

Medical AI systems also face challenges in high-resource languages such as English (Lee et al., 2024). Factual recalls alone are not sufficient for medical reasoning. It requires understanding the medical terminologies, logical inference, and the ability to capture clinical information (Singhal et al., 2025; Nori et al., 2023; Thapa et al., 2024). These challenges are severe in low-resource languages. Ideally, models should handle the same reasoning capabilities while working with limited training data, unfamiliar morphological structures, and scripts that differ fundamentally from their pre-training corpora. Each factor creates significant barriers and threatens the reliability of medical AI systems in underrepresented languages, including Urdu.

Urdu is the 11th most spoken language with over 246 million speakers worldwide (Lewis, 2009). The language exists in two forms: (1) standard Urdu uses the Perso-Arabic script (written right-to-left), and (2) Roman Urdu employs Latin characters for the same vocabulary. In digital spaces, Roman Urdu has become dominant (Ali et al., 2021; Ullah et al., 2024) as people text about health concerns, share medical information, and ask questions in this informal variant. To the best of our knowledge, there is no standardized way to eval-

[*] Both authors contributed equally to this work.

[1] https://www.ethnologue.com/insights/ethnologue200/

[2] https://github.com/AliFaheem/U-Mirage

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 453–460
March 28, 2026. ©2026 Association for Computational Linguistics

uate whether medical AI systems work in either form of Urdu. This absence of benchmarks prevents us from understanding how existing models perform and whether established techniques can bridge the performance gap.

Chain-of-Thought (CoT) prompting offers a potential solution (Wei et al., 2022; Kojima et al., 2022). By asking models to show their reasoning process step-by-step, CoT has improved performance on English medical tasks (Singhal et al., 2025; Thapa et al., 2024). But will this approach transfer to low-resource languages? Urdu appears infrequently in training data. Its morphological structure differs from English. The Perso-Arabic script bears little resemblance to the Latin text that dominates pre-training corpora. These differences raise fundamental questions about whether techniques that work in English will help or even harm performance in other languages.

Chain-of-Thought (CoT) prompting offers a potential solution (Wei et al., 2022; Kojima et al., 2022). By asking models to show their reasoning process step-by-step, CoT has improved performance on English medical tasks (Singhal et al., 2025; Thapa et al., 2024). However, whether this approach transfers to low-resource languages, including Urdu, remains unexplored. Several factors complicate this transfer, including (1) Urdu appears infrequently in training data compared to high-resource languages, (2) its morphological structure differs from English, and (3) the Perso-Arabic script bears little resemblance to the Latin text that dominates pre-training corpora. These limitations raise questions about whether techniques that work in English will help in other languages. Answering these questions requires creating an appropriate benchmark for Urdu.

Evaluating medical LLMs in Urdu faces various challenges. Most English medical terms lack standard Urdu translations, so speakers naturally incorporate English medical vocabulary into Urdu sentences. Roman Urdu has no official spelling rules, leading to variation even in how people write the same word. The script difference between Perso-Arabic Urdu and Latin-based training data may fundamentally affect how models process the language. These challenges create critical knowledge gaps about cross-lingual performance degradation, the effectiveness of prompting strategies in low-resource settings, and the role of model characteristics in multilingual medical reasoning.

We investigate three research questions:

RQ1: How do state-of-the-art LLMs perform on medical reasoning tasks when transitioning from English to Urdu and Roman Urdu?

RQ2: Can Chain-of-Thought prompting mitigate performance degradation in medical reasoning for these underrepresented languages?

RQ3: How do model architecture, size, and quantization affect cross-lingual medical reasoning capabilities?

To address these questions, we make three main contributions:

1. We introduce U-MIRAGE, the first medical question-answering benchmark for Urdu and Roman Urdu.

2. We evaluate six LLMs for English, Urdu, and Roman Urdu, using two strategies (1) zero-shot and (2) Chain-of-Thought prompting.

3. We also provide a systematic analysis of how model architecture, size, and quantization affect cross-lingual medical reasoning.

Our findings reveal that medical knowledge does not transfer uniformly across languages. This work provides benchmarks for Urdu and Roman Urdu medical QA, hence establishing the capacity requirements for effective cross-lingual medical reasoning.

2 Related Work

Large Language Models (LLMs) have demonstrated strong reasoning capabilities when guided by CoT prompting, which encourages models to decompose complex problems into interpretable intermediate steps. (Wei et al., 2022) reported substantial gains on complex mathematical reasoning tasks, while (Kojima et al., 2022) showed that even zero-shot CoT prompting can create structured reasoning without manually curated exemplars. The analysis revealed that CoT efficacy is highly sensitive to model scale and architecture, with smaller models exhibiting inconsistent or unstable improvements.

Recent work has shown potential progress in Medical AI for high-resource languages, where LLMs such as *Med-PaLM 2* have achieved medical expert-level accuracy on medical queries (Singhal et al., 2025; Nori et al., 2023). The standard benchmark datasets, including *MedQA*, *PubMedQA*, and *MedMCQA*, have been widely used to evaluate English medical question answering tasks (Jin et al., 2019; Pal et al., 2022). However, models show uneven performance across different question types. (Thapa et al., 2024) reported a 16.8% gap between factual recall and reasoning tasks, underscoring the need for capabilities beyond memorization.

Low-resource languages further increase the performance gap due to their limited representation in their training data. Multilingual models like mT5 and BLOOM perform 15% to 40% worse on low-resource reasoning tasks as compared to English (Qiu et al., 2024; Tahir et al., 2025). Research shows that reasoning accuracy is consistently lower than factual recall across almost all low-resource conditions (Nazi and Peng, 2024), underscoring the need for native-language examples, guided prompting strategies, and tailored evaluation methods.

Urdu, despite its global significance, is underrepresented in medical NLP (Arif et al., 2024). Although there is work on transliteration and translation between standard Urdu and Roman Urdu (Butt et al., 2025; Faheem et al., 2025), the medical reasoning abilities in these variants remain unexplored.

Our study addresses these gaps by introducing parallel medical QA datasets for Urdu and Roman Urdu, evaluating six LLMs across both linguistic variants, and analyzing how architectural factors, quantization levels, and cross-lingual transfer influence performance in low-resource medical NLP.

3 U-MIRAGE

We introduce U-MIRAGE, a medical question-answering dataset for Urdu and Roman Urdu based on the English MIRAGE benchmark (Xiong et al., 2024). The dataset contains 7,663 multiple-choice questions translated into both Urdu and Roman Urdu. We maintain alignment with the original English questions while preserving medical accuracy and naturalness. The MIRAGE aggregates questions from five established medical benchmarks (dataset: instances), i.e., (1) MedQA: 1,273, (2) MedMCQA: 4,183, (3) PubMedQA: 500, (4) BioASQ: 618, and (5) the medical subset of MMLU: 1,089. Each question presents four answer options with one correct choice.

3.1 Translation Protocols

Translating medical content into low-resource languages presents two main challenges: (1) semantic distortion and (2) terminology gaps. Direct word-for-word translation often produces clinically incorrect phrases that diverge from how medical professionals and patients actually communicate in Urdu. We address this challenge through code-mixing, reflecting how Indo-Pak healthcare practitioners naturally combine English medical terms with Urdu discourse. Our translation protocol retains English medical terms when Urdu equivalents are ambiguous or non-standard, therefore, prioritizing clinical accuracy over literal translation. For instance, we preserve "panic attack" and "allergy test" rather than using literal translations like *"ghabrahat ka hamla"* or *"hassasiyat ka imtehan"*, which sound unnatural and may cause confusion. This approach aligns with research on code-mixed NLP in specialized domains (Winata et al., 2023; Ahuja et al., 2023). Table 1 provides a few examples of the code-mixing approach.

English Term	**Literal Translation**
Panic attack	*Ghabrahat ka hamla*
Allergy test	*Hassasiyat ka imtehan*
Heart disease	*Dil ki bemari*
Blood pressure	*Khoon ka dabao*

Table 1: Code-mixing approach: preserving English medical terms rather than using literal Urdu translations.

We created two variants: MIRAGE_ru (Roman Urdu, using Latin script) and MIRAGE_ur (standard Urdu, using Perso-Arabic script). Combined with the original English MIRAGE dataset, this yields U-MIRAGE, a trilingual benchmark containing $7,663$ parallel questions across all three language variants.

3.2 Validation

To verify that translation preserved medical meaning, we conducted backtranslation experiments on 500 randomly sampled questions. We translated English questions into Roman Urdu, then back into English, and measured ROUGE overlap between the original and back-translated versions. High ROUGE scores indicate successful semantic preservation.

As shown in Table 2, the ROUGE-L F1 score of 0.75 confirms strong semantic preservation, demonstrating that the translation process maintains medical accuracy. We also measured lexical overlap between the English and Roman Urdu versions to quantify the extent to which medical terminology was preserved versus translated. Since both use the

Metric	Precision	Recall	F1
ROUGE-1	0.82	0.80	0.81
ROUGE-2	0.61	0.60	0.60
ROUGE-L	0.76	0.74	0.75

Table 2: Backtranslation validation using English to Roman Urdu to English (EN to RU to EN) with n=500.

Latin script, ROUGE can directly compare token overlap. Table 3 shows results by source dataset.

Dataset	ROUGE-1	ROUGE-2	ROUGE-L
MedQA	0.27	0.13	0.24
MedMCQA	0.28	0.13	0.25
PubMedQA	0.34	0.18	0.28
BioASQ	0.39	0.17	0.36
MMLU	0.24	0.10	0.21

Table 3: Lexical overlap between English and Roman Urdu by source dataset.

BioASQ shows the highest overlap (ROUGE-L: 0.36), while MMLU shows the lowest (0.21), reflecting differences in how technical terminology appears across medical subdomains.

4 Methodology

In this section, we will discuss our methodology, including model selection and prompting techniques.

4.1 Models

We evaluate six large language models spanning different architectures, sizes, and accessibility levels. Our selection includes three proprietary models, including GPT-5-Nano and Google Gemini 2.0 Flash Lite. Three open-source models (1) LLaMA 3.2, (2) Gemma 2, and (3) Qwen 2.5 are also used in our work. This combination allows us to assess state-of-the-art performance while examining how architectural choices and model capacity affect cross-lingual medical reasoning. The proprietary models represent the current best performance on medical QA tasks. The open-source models range from 0.5B to 3B parameters, enabling us to analyze the capacity requirements for multilingual medical reasoning. For several models, we test both full-precision and quantized variants to understand how compression affects performance in low-resource languages.

4.2 Prompting Strategies

We test two prompting approaches (zero-shot and CoT) across all models and languages. Zero-shot prompting provides only the question and answer options, measuring how well medical knowledge transfers across languages without additional guidance. Chain-of-Thought prompting examines models to explain their reasoning step-by-step before selecting an answer, testing whether explicit reasoning helps overcome linguistic barriers. Each model processes the complete dataset in each language variant with both prompting strategies. This yields six evaluation conditions per model (3 languages × 2 prompting strategies), allowing systematic comparison of how language choice and prompting approach interact with model characteristics. Figure 1 shows example prompts for both of our strategies.

We evaluate each model on U-MIRAGE across three languages, i.e., English, Roman Urdu, and Urdu. To ensure consistency, all experiments use a low temperature setting (temperature=0.1) to produce near-deterministic outputs. We measure performance using exact-match accuracy, where a model receives credit only if it selects the correct answer option.

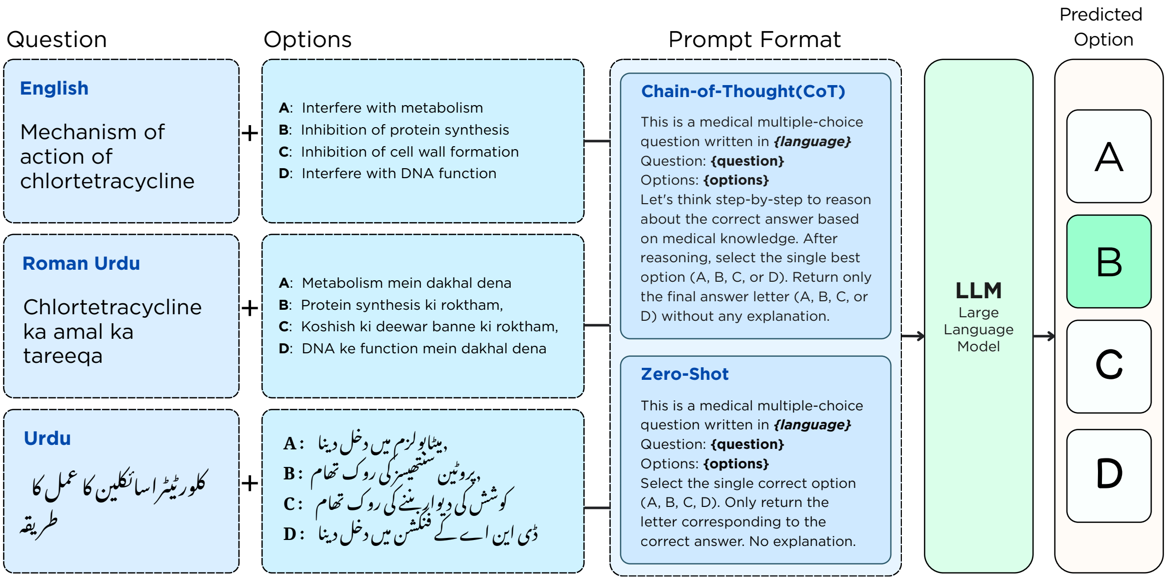
5 Results and Discussion

Table 4 presents accuracy results for all models across languages and prompting strategies. We organize our findings around three main observations.

5.1 Cross-Lingual Performance

All models show substantial accuracy drops when moving from English to Urdu variants. GPT-5-Nano achieves 80.77% on English but falls to 74.61% on Roman Urdu and 73.00% on Urdu, drops of 6.16% and 7.77%, respectively. Gemini 2.0 Flash Lite shows even larger degradation: from 71.78% in English to 65.83% in Roman Urdu (5.95%) and 68.09% in Urdu (3.69%). This pattern holds across all models in zero-shot settings, with performance losses ranging from 6% to 11%. The consistency of this degradation across different architectures suggests fundamental challenges in transferring medical knowledge to low-resource languages, rather than model-specific limitations.

Roman Urdu generally outperforms standard Urdu script. Across the models, Roman Urdu shows improvements of 0.9% to 1.6% over Urdu. The one exception is Gemini 2.0 Flash Lite, where



A: Interfere with metabolism
B: Inhibition of protein synthesis
C: Inhibition of cell wall formation
D: Interfere with DNA function

A: Metabolism mein dakhal dena
B: Protein synthesis ki roktham,
C: Koshish ki deewar banne ki roktham,
D: DNA ke function mein dakhal dena

A: میٹابولزم میں دخل دینا,
B: پروٹین, سنتھیسز کی روک تھام
C: کوشش کی دیوار بننے کی روک تھام
D: ڈی این اے کے فنکشن میں دخل دینا

This is a medical multiple-choice question written in ***{language}*** Question: **{question}** Options: **{options}** Let's think step-by-step to reason about the correct answer based on medical knowledge. After reasoning, select the single best option (A, B, C, or D). Return only the final answer letter (A, B, C, or D) without any explanation.

This is a medical multiple-choice question written in ***{language}*** Question: **{question}** Options: **{options}** Select the single correct option (A, B, C, D). Only return the letter corresponding to the correct answer. No explanation.

Figure 1: Example prompts for zero-shot and Chain-of-Thought strategies.

Model	English		Roman Urdu		Urdu	
	Zero-Shot	CoT	Zero-Shot	CoT	Zero-Shot	CoT
GPT-5-Nano	**80.77**	**78.70**	**74.61**	**75.01**	**73.00**	**70.25**
Gemini 2.0 Flash Lite	71.78	77.79	65.83	74.13	68.09	70.15
Qwen 2.5 0.5B (quantized)	25.92	30.31	23.78	28.69	23.04	27.55
Gemma 2 2B (quantized)	42.03	43.29	32.46	34.55	31.24	33.67
LLaMA 3.2 1B (quantized)	18.85	38.99	19.82	30.54	18.83	29.48
LLaMA 3.2 1B	45.55	45.19	34.53	38.52	33.66	32.66

Table 4: Accuracy (%) across models, languages, and prompting strategies on U-MIRAGE.

Urdu performs 2.26% better than Roman Urdu. This advantage for Roman Urdu likely reflects its orthographic similarity to the Latin-script text dominating pre-training corpora.

5.2 Chain-of-Thought Effects

CoT prompting produces inconsistent effects across model architectures. Gemini 2.0 Flash Lite benefits substantially: CoT improves English performance by 6.01% (from 71.78% to 77.79%), Roman Urdu by 8.30% (from 65.83% to 74.13%), and Urdu by 2.06% (from 68.09% to 70.15%). The quantized LLaMA 3.2 1B shows even larger gains with 20.14% in English and approximately 10.7% in both Urdu and Roman Urdu variants. In contrast, GPT-5-Nano performs worse with CoT prompting. English accuracy drops 2.07% (from 80.77% to 78.70%), and Urdu drops 2.75 (from 73.00% to 70.25%). Only Roman Urdu shows a marginal improvement of 0.40%.

Figure 2 visualizes these patterns through radar charts comparing CoT and zero-shot performance across languages for each model. Figure 3 shows the magnitude of changes, where Gemini displays predominantly positive shifts while GPT-5-Nano shows mixed results.

These results contradict the assumption that CoT prompting universally improves reasoning. Instead, its effectiveness depends critically on model architecture.

5.3 Quantization Impact

Quantized models with fewer than 2B parameters fail catastrophically in medical QA, particularly in low-resource languages. Qwen 2.5 0.5B (quantized) achieves only 25.92% on English, which is

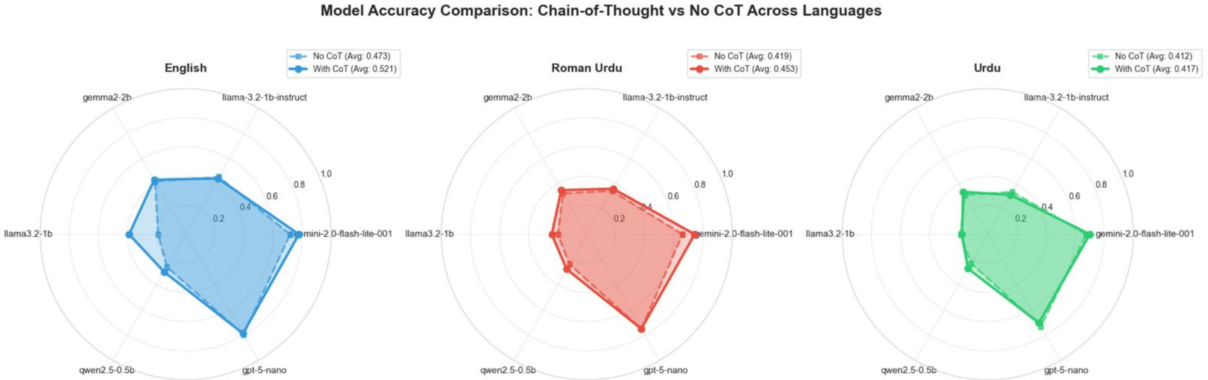


Figure 2: Radar charts comparing zero-shot and CoT performance across languages for each model.

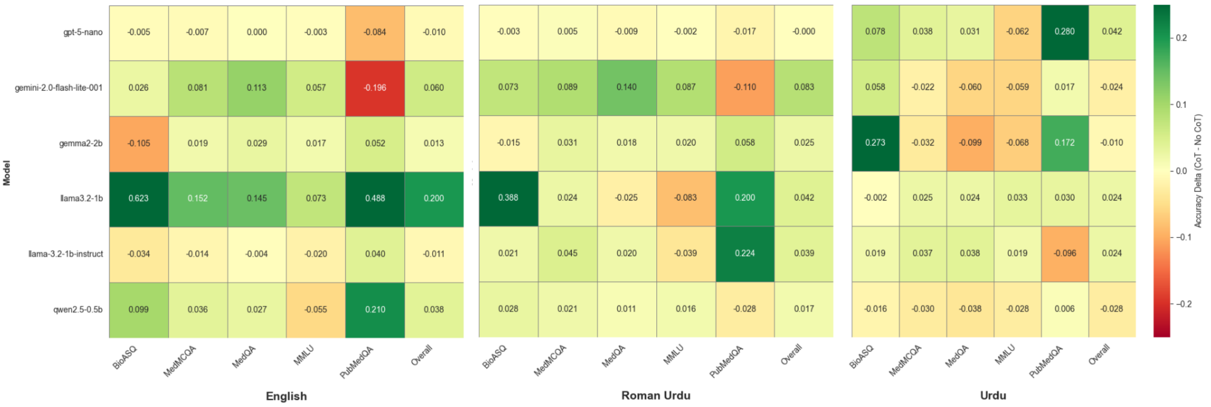


Figure 3: Heatmap showing percentage changes from zero-shot to CoT prompting across models and languages.

barely above random chance for four-option multiple choice. Performance worsens in Urdu variants: 23.78% in Roman Urdu and 23.04% in Urdu. Even with CoT prompting, accuracy remains below 31% across all languages.

Comparing full-precision and quantized versions of LLaMA 3.2 1B reveals the severity of quantization penalties. The full-precision model achieves 45.55% on English, while the quantized version achieves only 18.85%, resulting in a gap of 26.70%. While CoT prompting recovers 20.14% for the quantized model in English, performance still lags far behind the full-precision baseline. These results establish a practical threshold: models need at least 2B parameters in full precision to perform medical reasoning in low-resource languages. Smaller quantized models fail regardless of prompting strategy, making them unsuitable for medical applications in these contexts.

Our findings reveal critical limitations in current approaches to multilingual medical AI. Medical knowledge does not transfer uniformly across languages despite being theoretically language-independent. The 6% to 11% drop in performance when moving to Urdu variants suggests that current models rely heavily on surface-level patterns in English medical text rather than deeper conceptual understanding. CoT prompting is not a universal solution. While it helps some models substantially, it harms others. This architecture dependence means practitioners cannot simply apply CoT prompting to any model and expect improvement; they must validate its effectiveness for their chosen model. Model compression through quantization creates unacceptable trade-offs for medical reasoning in low-resource languages. Quantized models with fewer than 2B parameters perform at near-random levels, making them dangerous for medical applications where accuracy matters.

The consistent advantage of Roman Urdu over standard Urdu script points to a practical consideration: orthographic similarity to pre-training data may matter more than linguistic proximity. This suggests that for languages with multiple scripts, choosing the variant closer to Latin characters may improve performance with current models. These results have direct implications for deploying medical AI in linguistically diverse regions. Organizations cannot assume that models that perform well in English will perform adequately in other languages. They must validate performance in target languages, carefully select model architectures and prompting strategies, and ensure sufficient model capacity to avoid catastrophic failure.

6 Conclusion

We introduced U-MIRAGE, the first medical question-answering benchmark. U-MIRAGE comprises 7,663 question-and-answer pairs across three languages, including Urdu, Roman Urdu, and English. We evaluate six LLMs and reveal three key findings. (1) All models show a 6% to 11% drop in performance when moving from English to Urdu, indicating that medical knowledge does not transfer uniformly across languages. (2) Chain-of-Thought prompting produces architecture-dependent effects; therefore, some models improved by 8% to 20% while some degraded by up to 3%. (3) Quantized models (2B parameters) achieved near random accuracy regardless of various prompting strategies. Notably, Roman Urdu consistently outperforms standard Urdu, suggesting that orthographic alignment with pre-training data matters more than linguistic proximity. Future work should explore the pre-training on domain-specific medical data and investigate whether multilingual pre-training improves the cross-lingual reasoning capabilities.

Limitations

U-MIRAGE relies on translated content from the English medical benchmark dataset, potentially limiting its ability to capture native linguistic patterns in Urdu. Our evaluation covers six LLMs due to computational constraints, precluding broader architectural comparisons. Additionally, our assessment focuses on the precision of multiple-choice responses, without examining the quality of the explanation or clinical safety.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, and 1 others. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267.

Hazrat Ali, Khalid Iqbal, Ghulam Mujtaba, Ahmad Fayyaz, Mohammad Farhad Bulbul, Fazal Wahab Karam, and Ali Zahir. 2021. Urdu text in natural scene images: a new dataset and preliminary text detection. *PeerJ Computer Science*, 7:e717.

Samee Arif, Abdul Hameed Azeemi, Agha Ali Raza, and Awais Athar. 2024. Generalists vs. Specialists: Evaluating Large Language Models for Urdu. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7263–7280.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Umer Butt, Stalin Varanasi, and Günter Neumann. 2025. Low-Resource Transliteration for Roman-Urdu and Urdu Using Transformer-Based Models. In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*, pages 144–153.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, and 1 others. 2023. PaLM: Scaling Language Modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113.

A. Faheem, F. Ullah, U. Azam, and 1 others. 2025. Part of speech (POS) tagging in Roman Urdu: datasets and models. *Language Resources & Evaluation*, 59:4285–4312.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *arXiv preprint arXiv:2004.09095*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2024. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

M Paul Lewis. 2009. *Ethnologue: Languages of the world*. SIL international Dallas, TX.

Zabir Al Nazi and Wei Peng. 2024. Large Language Models in Healthcare and the Medical Domain: A Review. *Informatics*, 11(3):57.

Harsha Nori, Nicholas King, Scott M McKinney, Daniel Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375*.

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.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards Building Multilingual Language Models for Medicine. *Nature Communications*, 15(1):8384.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Munief Hassan Tahir, Sana Shams, Layba Fiaz, Farah Adeeba, and Sarmad Hussain. 2025. Benchmarking the Performance of Pre-trained LLMs across Urdu NLP Tasks. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 17–34.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, and 1 others. 2024. Disentangling Reasoning and Knowledge in Medical Large Language Models. *arXiv preprint arXiv:2505.11462*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Faizad Ullah, Ali Faheem, Ubaid Azam, Muhammad Sohaib Ayub, Faisal Kamiran, and Asim Karim. 2024. Detecting cybercrimes in accordance with Pakistani law: Dataset and evaluation using PLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4717–4728, Torino, Italia. ELRA and ICCL.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in neural information processing systems*, 35:24824–24837.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, and 1 others. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251. Association for Computational Linguistics.