MedHalwasa: Quantify & Analyze Factual Hallucinations in Large Language Models For Arabic Medical Data

Anonymous ACL submission

Abstract

Hallucination in medical text generation poses critical risks, especially when large language models (LLMs) produce factually incorrect information. Such behavior, particularly when occurring at scale, can affect the quality of clinical decision-making and compromise patient safety. Although this issue has been studied in English, it remains largely unexplored in Arabic. We introduce ***MedHalwasa***, the first Arabic dataset to quantify and analyze hallucination in Arabic medical fact generation. The name *MedHalwasa* is derived from "Medical Halwasa," where "Halwasa" denotes hallucination in Arabic. Using nine different LLMs, we generate and evaluate 9,000 Arabic medical facts, annotating them with automatic factuality annotations. To support future research, we detail a systematic and reproducible data generation and annotation framework that can be extended to study other LLMs and domains. Our study enables the first systematic analysis of hallucinations in Arabic medical contexts and offers key insights to inform the selection of reliable LLMs for Arabic healthcare applications. The dataset is publicly accessible to facilitate future research[1].

1 Introduction

The rise of foundation models has marked a turning point in the field of Artificial Intelligence (AI), unlocking new possibilities across a broad spectrum of applications. Among these, Large Language Models (LLMs) have stood out for their ability to perform a wide array of natural language processing (NLP) tasks such as summarization, translation, question answering, and text generation with impressive fluency and coherence. By learning patterns from massive amounts of general and domain-specific text, LLMs have shown not only linguistic competence but also surprising adaptability, making them increasingly appealing for deployment in sensitive and high-impact areas such as healthcare.

Despite their remarkable capabilities, LLMs are prone to a critical limitation commonly referred to as hallucination, the generation of information that appears fluent and convincing yet is factually incorrect or unverifiable (Rawte et al., 2023)(Ji et al., 2023). Previous research (Huang et al., 2025b) has broadly classified hallucinations into two main categories: factuality hallucination, where the generated content conflicts with known or verifiable real-world facts, and faithfulness hallucination, where the output deviates from the given input or exhibits internal inconsistency. These distinctions underscore the complexity of the hallucination problem and highlight the pressing need for robust mechanisms to detect, quantify, and mitigate hallucinations.

In general-purpose applications, such outputs may be relatively harmless; however, in medical contexts, hallucinations carry far more serious implications. They can misguide clinical decision-making, misinform patients, and ultimately pose risks to patient safety. Hallucinated medical content may include incorrect dosages, fabricated conditions, or misleading interpretations of diagnostic criteria, all of which have the potential to cause real-world harm (Kim et al., 2025).

Although numerous studies have investigated the risks of hallucination of LLMs in the medical domain for English (Agarwal et al., 2024) (Ahmad et al., 2023) (Ziaei and Schmidgall, 2023a), there remains a limited understanding of how this problem manifests in medium and low-resource languages such as Arabic, where the impact may be equally, if not more, concerning due to the scarcity of language-specific resources and benchmarks (Silly et al., 2025).

The need for deeper investigation into hallucinations in LLMs within the context of Arabic medical information is underscored by both the linguistic

[1] Available upon request

complexity of Arabic and the critical importance of accurate medical communication. Addressing hallucinations in this setting is not only essential for enhancing the reliability and trustworthiness of LLMs but also holds significant implications for deploying such models in regions with limited medical infrastructures and low-resource healthcare systems. Therefore, the motivation behind our work is twofold: (1) to enable a rigorous and reproducible evaluation of the factuality and reliability of LLMs when generating medical content in Arabic, and (2) to pave the way for developing robust techniques to detect and mitigate hallucinations in Arabic medical LLMs.

This paper aims to bridge the current understanding gap by conducting a comprehensive study on hallucinations in the generation of Arabic medical texts. Using a diverse set of LLMs with varying architectures, sizes, and training characteristics, we seek to answer the following research questions.
I) To what extent can we rely on the factual accuracy of medical content generated by LLMs in Arabic?
II) Do specific model attributes (e.g., model size, training language mix, or reasoning capabilities) play a consistent role in mitigating factual hallucinations?
III) Which medical domains are most susceptible to factual hallucinations, and do certain models repeatedly fail on specific medical concepts within Arabic texts?

2 Related Work

Hallucinations in Medical Text Generation. Hallucinations in medical text generation present unique challenges compared to those in general domains (Kim et al., 2025). They often appear in critical tasks such as diagnosing conditions, recommending treatments, or interpreting laboratory results, where even small inaccuracies can have serious consequences for patient care. What makes them particularly concerning is that these errors often sound plausible and include specialized medical terms, which can make them difficult to spot without expert knowledge. As healthcare systems begin to integrate AI more deeply into clinical workflows, the risk of such hallucinations going unnoticed becomes even more pressing.

A growing body of research has examined hallucinations in large language models within the medical domain, emphasizing the serious risks posed by factually inaccurate outputs that can undermine clinical decision-making and compromise patient safety (Agarwal et al., 2024; Ahmad et al., 2023; Ziaei and Schmidgall, 2023b). In response, several benchmarks have been proposed to systematically detect and evaluate hallucinations in medical LLMs on a variety of task types, input modalities, and evaluation protocols (Agarwal et al., 2024).

One such benchmark is Med-HALT (Umapathi et al., 2023), which evaluates hallucinations in LLMs through two categories of tests: Reasoning Hallucination Tests (RHTs) and Memory Hallucination Tests (MHTs), designed to assess models' capabilities in problem-solving and information recall. Meanwhile, HALT-MedVQA (Wu et al., 2024) investigates hallucinations in large vision-language models (LLVMs) and introduces a dataset of medical images paired with question-answer sets. HALT-MedVQA focuses on three core tasks: FAKE Question, None of the Above (NOTA), and Image SWAP, and evaluates model performance primarily using LLaVA-based architectures (Liu et al., 2023).

In the context of non-English medical data, the Chinese Medical Hallucination Evaluation Benchmark (CMHE) (Dou et al., 2024) offers a comprehensive evaluation to understand and mitigate hallucination issues in Chinese medical LLMs. It identifies a crucial issue called "snowballing hallucination", where LLMs generate more errors when encountering initial misinformation. The CMHE benchmark features three primary tasks: detecting hallucinations, diagnosing diseases in complex scenarios, and explaining medical concepts, using specific medical glossaries such as ICD-10 and MeSH for robust evaluation.

Several other benchmarks have also contributed to this emerging field by providing structured evaluations of hallucinations in both LLMs and vision-language models (Gu et al., 2024; Chen et al., 2024; Manes et al., 2024; Addlesee, 2024; Hegselmann et al., 2024; Zuo and Jiang, 2024). Together, these efforts offer important tools and insights for understanding and mitigating hallucination risks in medical AI systems.

Hallucinations in Arabic Text Generation. Despite the growing interest in hallucinations produced by LLMs in high-resource languages, relatively few studies have investigated this phenomenon in Arabic, a morphologically rich and syntactically complex language that presents dis-

tinct challenges for language modeling. Halwasa (Mubarak et al., 2024) introduced the first Arabic data set specifically designed to quantify and analyze hallucinations in LLM output. This work focuses on open-ended text generation and evaluates two OpenAI models, with generated outputs manually annotated for factual accuracy, correctness, and linguistic quality. While Halwasa targets sentence-level hallucinations in Arabic, HalluVerse25 (Abdaljalil et al., 2025) offers a multilingual benchmark that includes fine-grained hallucination categories across English, Arabic, and Turkish.

Hallucinations in Arabic Medical Text Generation. To the best of our knowledge, the intersection of medical hallucinations and Arabic text generation remains largely underexplored, with no existing datasets or benchmarks dedicated to evaluating factual consistency in this critical setting. Moreover, existing research shows that while LLMs may be capable of detecting hallucinations, this ability does not necessarily prevent them from generating hallucinated content in the first place (Dou et al., 2024). In light of these gaps, we introduce *MedHalwasa*, a dataset designed to evaluate the factuality of Arabic medical content generated through open-ended generation. Unlike previous work that tests the ability of LLMs to detect hallucinations, our focus is on assessing the likelihood that the models generate hallucinated medical information in the first place.

3 Data Generation

To assess the factuality of the Arabic medical content generated by large language models, we design *MedHalwasa* data generation and evaluation framework in three main stages as shown in Figure 1. These stages are further decomposed into five key steps: (1) generation of medical vocabulary, (2) LLM selection strategy, (3) data generation protocol, (4) factuality evaluation, and (5) results analysis.

Medical Vocabulary Generation. To ensure broad coverage across various medical subdomains, we first generate a list of Arabic medical terms that reflect the most frequently queried topics by Arabic-speaking users. We employ structured prompting with ChatGPT to obtain a list of 200 Arabic medical terms. The prompt used is illustrated in Figure 2.

Large Language Model Selection Strategy. We choose a diverse set of LLMs to reflect different behaviors in Arabic medical fact generation. Our selection covers key aspects such as model size, accessibility (open-source vs. closed-source), language scope (bilingual vs. multilingual), and specialization (reasoning-focused vs. general-purpose). The models we evaluate include GPT-4o (OpenAI et al., 2024b), o1 (OpenAI et al., 2024a), Qwen2.5-Max (Team, 2024), DeepSeek-R1 (DeepSeek-AI et al., 2025), jais-adapted-70b (Inception, 2024), Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), BiMediX2-8B-hf (Mullappilly et al., 2024), and Fanar (Team et al., 2025).

Data Generation Protocol. With the defined vocabulary and models, we implement a unified data generation protocol to ensure consistency between models. Following (Mubarak et al., 2024), each model is prompted to generate five verifiable factual sentences per Arabic medical term, yielding a total of 1,000 factual statements per model. We refer to this aggregated dataset of 9,000 Arabic medical factual sentences as the *MedHalwasa* dataset. The prompt structure used across all models is shown in Figure 3.

Due to differences in model availability and behavior, we use three distinct interaction methods: (1) API calls for GPT-4o, o1, Fanar, Llama-3.1-8B-Instruct, Llama-3.1-8B-Instruct, and BiMediX2-8B-hf (2) local hosting for Jais-70b, and (3) web-based prompting through Macro Recorder for models with undesirable behavior when called via the API such as Qwen2.5-Max and DeepSeek-R1. Some of the models exhibit language inconsistencies, occasionally mixing Arabic with English or Chinese text. Others fail to respond or generate improperly formatted outputs (e.g., a paragraph instead of separate facts). To address these inconsistencies, we generate multiple runs when needed and select the most suitable outputs from among them.

Factuality Assessment. To evaluate the factual accuracy of the generated Arabic medical sentences in the *MedHalwasa* dataset, we employ two automatic factuality assessment frameworks, the OpenFactCheck framework (Iqbal et al., 2024) and the MedScore (Huang et al., 2025a), both are designed through a two-step decompose-then-verify approach specifically suitable for free-form an-

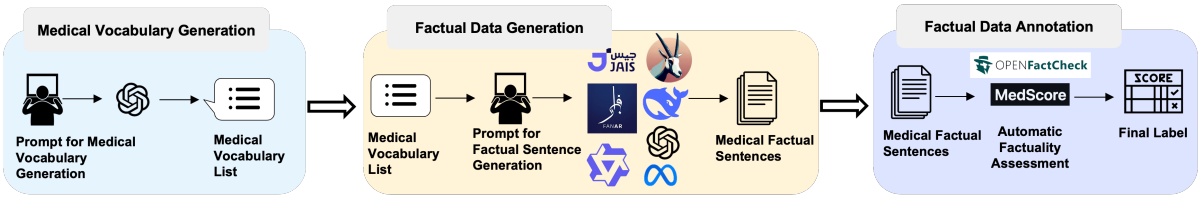

Figure 1: Overview of the *MedHalwasa* dataset construction and evaluation pipeline. It comprises three main stages: (1) Arabic medical vocabulary generation using ChatGPT, (2) factual data generation, where multiple LLMs produce Arabic medical sentences based on the vocabulary list, and (3) data annotation, where automatic factuality assessment frameworks assign factuality labels to the generated sentences.

> **Provide me a list of 200 Arabic medical terms queried by individuals seeking medical information or checking medical presuppositions that are frequently used in hospitals and online platforms. list the terms as follows :**
>
> **Arabic medical term , English translation of Arabic medical term**

Figure 2: Prompt used in Chat-GPT for Arabic medical terms generation.

> **Give exactly FIVE Arabic complete and diverse factual sentences having the following term: {term}. These sentences should have facts that can be checked and verified. Write the sentences separated by a new line without translation and without numbering.**

Figure 3: Prompt used in the nine selected LLM models to generate the *MedHalwasa* dataset.

swers.

For OpenFactCheck, we developed a custom script to process all 9,000 generated sentences. The framework first uses a Claim Processor to decompose each sentence into a list of atomic, check-worthy claims. These claims are then passed to a Retriever, which collects supporting evidence from external sources, including Google Search, Wikipedia, and GPT-based retrieval. Finally, a Verifier evaluates each claim against the retrieved evidence to assign a binary factuality label: True or False.

MedScore, on the other hand, is tailored to the medical domain and focuses on decomposing medical responses into condition-aware, clinically relevant facts. It verifies each claim using the internal knowledge of the general-purpose model and medical-specialized model to determine its truthfulness. MedScore outputs a continuous score between 0 and 1, representing the proportion of true claims in a given sentence.

It is worth noting that OpenFactCheck produces a strict binary judgment (1: True, 0: False), making it a more conservative evaluator compared to MedScore, which offers a more nuanced measure

Arabic Medical Term	English Medical Term
التهاب الحلق	Sore throat
التهاب اللوزتين	Tonsillitis
ارتفاع ضغط الدم	Hypertension
انخفاض ضغط الدم	Hypotension
السكري	Diabetes
السكتة الدماغية	Stroke
النوبة القلبية	Heart attack
انسداد الشرايين	Artery blockage

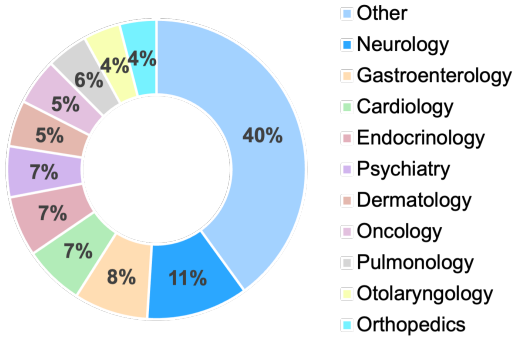

Figure 4: Top: The table on the top presents a subset of Arabic medical terms along with their corresponding English translations. Bottom: A donut chart showing the distribution of terms across the ten most frequent medical domains in the dataset; the remaining 25 domains are grouped under "Other".

of factuality.

4 Results

4.1 Medical Vocabulary and Data Generation

A sample from the generated list of 200 Arabic medical vocabulary is presented in Figure 4. By carefully engineering the prompt, we ensured that the extracted vocabulary encompasses a broad range of commonly queried medical terms, spanning symptoms, conditions, treatments, and diag-

Medical Term	Index	Generated Sentences
التهاب الحلق	1	يعتبر التهاب الحلق من الأمراض الشائعة التي تصيب الإنسان، وهو يحدث نتيجة لعدوى فيروسية أو بكتيرية في منطقة البلعوم والحنجرة.
التهاب الحلق	1	يمكن أن يتسبب التهاب الحلق بأعراض مثل الألم وصعوبة البلع وارتفاع درجة الحرارة والصداع والتعب العام.
التهاب الحلق	1	يتم علاج التهاب الحلق عادة باستخدام المضادات الحيوية إذا كان السبب بكتيريا، بينما تستخدم الأدوية المضادة للفيروسات لعلاج العدوى الفيروسية.
التهاب الحلق	1	يجب على الأشخاص الذين يعانون من أعراض حادة لالتهاب الحلق مراجعة الطبيب للحصول على التشخيص الصحيح والعلاج المناسب.
التهاب الحلق	1	ينصح بتناول السوائل الدافئة والغرغرة بالماء المالح لتخفيف الأعراض المصاحبة لالتهاب الحلق والحفاظ على ترطيب الفم والحلق.
التهاب اللوزتين	2	يمكن أن يحدث التهاب اللوزتين بسبب عدوى بكتيرية أو فيروسية، ويمكن أن يسبب أعراضًا مثل الألم في الحلق وصعوبة البلع والحمى وتضخم الغدد الليمفاوية في الرقبة.
التهاب اللوزتين	2	يتم علاج التهاب اللوزتين البكتيري عادة بالمضادات الحيوية، بينما يعالج الفيروسي بالراحة وتناول السوائل الدافئة والمسكنات لتخفيف الأعراض.

Figure 5: A sample from the *MedHalwasa* dataset showcasing generated Arabic medical factual sentences. Each row includes a medical term, its index (indicating the vocabulary group), and a corresponding factual sentence generated by an LLM. For each term, five different factual sentences are created to capture diverse factual expressions.

nostics frequently encountered in hospitals and online health platforms, corresponding to various medical domains. In total, there are 35 medical domains covered by the 200 medical terms. This comprehensive vocabulary base enables the generation of diverse and contextually rich factual statements in the *MedHalwasa* dataset, as illustrated in Figure 5.

4.2 Factual Accuracy of Generated Content

To assess the reliability of large language models in generating factual medical content in Arabic, we evaluate their factual accuracy. Table 1 presents a comparative evaluation of various large language models on the task of generating factual Arabic medical statements. We report performance using two factual evaluation frameworks: MedScore, which is specifically designed to assess medical factuality (Huang et al., 2025a), and OpenFactCheck, a domain-agnostic factuality framework applicable across various topics, including medicine (Iqbal et al., 2024). Although both metrics are designed for evaluating free-form text, they differ in granularity: MedScore assigns a continuous score between 0 and 1, allowing partial credit, whereas OpenFactCheck uses a binary judgment (0 or 1) for each statement. As a result, MedScore values tend to be higher on average, though both metrics show consistent relative rankings across models.

The table is sorted by MedScore values and divided into two groups: large-scale models and medium/small-scale models. Among the large models, Qwen2.5-Max achieves the highest MedScore (97.6%), demonstrating strong capability in generating factually accurate medical content in the Arabic language. This is followed by GPT-4o and o1 OpenAI models, both achieving a MedScore of 95.1%, with o1 distinguished by its reasoning capability. Deepseek-R1, a large open-source model (671B parameters), scores slightly lower (94.1%). However, this model activates a significantly smaller number of parameters at inference time. These results suggest that while reasoning capabilities are intended to improve factuality, in practice, they do not always translate to superior performance. Instead, models like Qwen2.5-Max and GPT-4o, although not explicitly labeled as reasoning models, demonstrate a stronger overall factual accuracy in the Arabic medical domain.

In the medium and small model category, Fanar-9B achieves the highest MedScore (92.3%), showcasing that a well-curated and culturally aligned bilingual (English-Arabic) model can compete effectively with significantly larger systems. Fanar is based on Google's Gemma-2-9B (Team et al., 2024) and has been continually pretrained with particular attention to Modern Standard Arabic (MSA) and various Arabic dialects. It also emphasizes alignment with Islamic values and Arab cultures, which may contribute to its superior factual grounding in Arabic content generation. jais-adapted-70b, a bilingual English-Arabic model adapted from Llama-2 (Touvron et al., 2023) , follows closely with a MedScore of 91.9%. Its strong Arabic proficiency, stemming from training on a massive high quality dataset of Arabic text, enables it to perform robustly despite being smaller than the large top-tier models.

LLaMA-3.3-70B-Instruct, a multilingual model that does not explicitly support Arabic, still achieves a moderate MedScore of 86.5%. While not optimized for Arabic nor the medical domain, its scale and general multilingual training allow for partial generalization to Arabic medical content, especially when compared to significantly smaller models.

BiMediX2-8B-hf, an English version of the BiMediX2 vision-language model built upon LLaMA-3.1 and trained on a large-scale bilingual and multimodal healthcare dataset, scores 76.8%, falling notably behind Fanar-9B despite being domain-specific. This suggests that while BiMediX2 benefits from domain alignment and bilingual pretraining, its multimodal architecture or reliance on the LLaMA-3.1 base may limit its factual accuracy in text-only generation tasks.

Model	Size	Type	MedScore	OpenFactCheck
Qwen2.5-Max	**325B**	**Multilingual/Open Source**	**97.6%**	**94.2%**
GPT-4o	–	Multilingual/Closed Source/VLM	95.1%	90.1%
o1	–	Multilingual/Closed Source/Reasoning/ VLM	95.1%	89.0%
Deepseek-R1	671B	Multilingual/Open Source/Reasoning	94.1%	85.7%
Fanar	9B	Bilingual/Open Source	92.3%	81.3%
jais-adapted-70b	70B	Bilingual/Open Source	91.9%	82.8%
Llama-3.3-70B-Instruct	70B	Multilingual/Open Source	86.5%	72.4%
BiMediX2-8B-hf	8B	Bilingual/Open Source/VLM	76.8%	57.5%
Llama-3.1-8B-Instruct	8B	Multilingual/Open Source	50.5%	40.9%

Table 1: Performance comparison of the factual accuracy of different LLMs in generating Arabic medical sentences, evaluated using MedScore and OpenFactCheck over 1,000 samples per model.

That said, BiMediX2-8B-hf demonstrates a significant improvement over its base model, LLaMA-3.1-8B-Instruct, which scores the lowest MedScore of 50.5%. This improvement highlights the value of task-specific fine-tuning on curated English-Arabic medical datasets. Although both are small models (8B), the contrast between them underscores the importance of domain adaptation, even within the same architecture family.

Turning to the OpenFactCheck results, we observe a consistent ranking pattern, though absolute scores are generally lower due to its binary scoring nature. Qwen2.5-Max again leads with 94.2%, followed by GPT-4o (90.1%), o1 (89.0%), and Deepseek-R1 (85.7%). Among smaller and medium models, Fanar (81.3%) and jais-adapted-70b (82.8%) retain strong performance, with the rest of the models exhibiting a gradual decline—culminating with Llama-3.1-8B-Instruct at 40.9%.

4.3 Cross-Model Failures in Medical Terms and Domains

While the previous analyses focus on model-specific performance, a cross-model investigation reveals consistent patterns of failure shared by multiple LLMs, particularly with respect to certain medical terms and domains. For each model, we analyzed the 1,000 generated Arabic medical facts to identify the top 10 domains and top 15 medical terms with the highest inaccuracy rates. We then aggregated these findings across all nine models, yielding insights from a total of 9,000 generated samples. The individual failure patterns for each model are presented in Figure 7, Figure 8, and Figure 9 in Appendix A, while the cross-model aggregation is summarized in Figure 6. These recurring inaccuracies suggest that certain medical concepts are systematically misrepresented, highlighting persistent challenges in Arabic medical understanding across current LLMs, regardless of model architecture or training approach.

Recurring Inaccurate Medical Terms. Among the most frequently misrepresented medical terms, leaky gut syndrome and optic neuritis stood out, appearing in the top inaccurate outputs of 5 out of 9 models (55.6%). Other recurring terms include glossitis, compression fractures, and prostatitis, which appeared in the top inaccurate terms of 3–4 models. Notably, even state-of-the-art models like GPT-4o and Qwen2.5-Max contributed to these inaccuracies, reinforcing the notion that term-

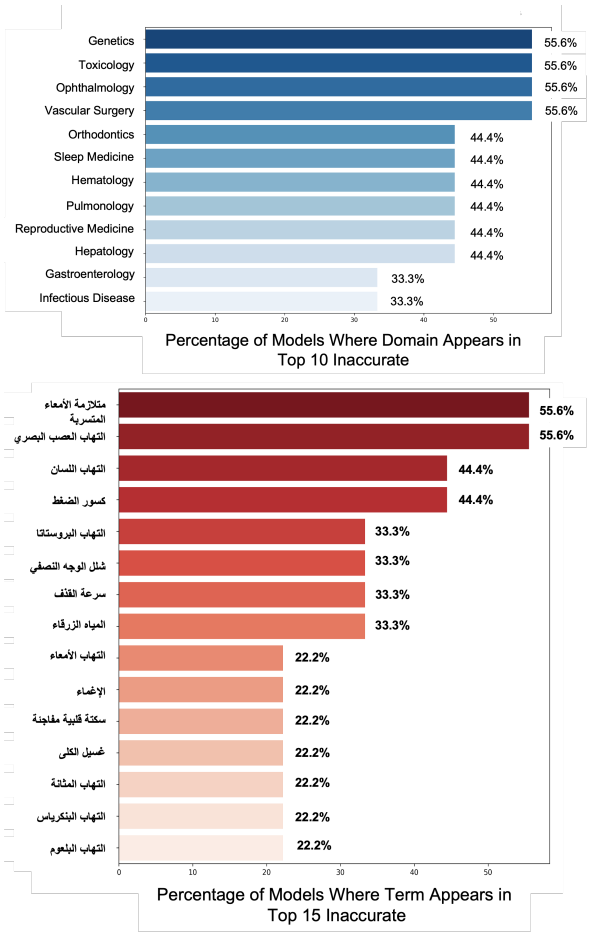

Figure 6: Aggregated analysis of common inaccuracies across all evaluated models. Top: highlights the medical domains that most frequently appear in the top inaccuracy lists across the nine models. Bottom: presents the most commonly occurring inaccurate medical terms in the generated content. These visualizations reveal consistent failure patterns that can inform future targeted model enhancements.

specific challenges are systemic rather than isolated.

Many of these terms reflect complex or less commonly encountered medical conditions, which may suffer from limited representation in Arabic-language medical corpora. Additionally, the morphological and syntactic complexity of Arabic may compound the difficulty in modeling precise clinical terms, leading to hallucinations or vague outputs in these areas.

Challenging Medical Domains. At the domain level, Genetics, Toxicology, Ophthalmology, and Vascular Surgery were the most frequently observed failure categories, each appearing in the top inaccurate domains of 5 models (55.6%). Close behind were Orthodontics, Sleep Medicine, Hematology, and Pulmonology, each flagged by 4 models. These domains often involve technical terminology, rapidly evolving research, and nuanced clinical context, making accurate generation especially challenging.

We hypothesize that these domain-level failures arise from two key factors: (1) data scarcity in high-quality Arabic medical texts for these domains, and (2) the models' inability to generalize reliably across specialized areas without explicit domain adaptation or retrieval-based augmentation.

5 Conclusion

In this work, we present *MedHalwasa*, the first comprehensive dataset generated for evaluating hallucinations in Arabic medical fact generation. Recognizing the critical implications of factual errors in clinical contexts, our study systematically examines the reliability of nine prominent large language models by generating and analyzing 9,000 factual Arabic medical statements. We propose a scalable and reproducible framework for data generation and automatic factual annotation, enabling consistent evaluation across models. Our dual metric approach, using both domain-specific (MedScore) and domain-agnostic (OpenFactCheck) evaluators, provides a nuanced understanding of model performance.

Through detailed analyses, we observe notable disparities in factual accuracy across model scales and configurations. Large-scale LLMs consistently outperform smaller and medium-sized models in Arabic medical fact generation, underscoring the impact of scale on factual reliability. Among the smaller and medium-scale models, those trained on high-quality Arabic corpora demonstrate stronger performance compared to general-purpose counterparts. Additionally, domain specialization plays a significant role, where a medically tuned model shows clear improvements over its base, general-purpose variant. Our cross-model examination also uncovers recurring inaccuracies tied to specific medical terms and domains, suggesting deeper limitations in current LLMs' medical reasoning and Arabic linguistic understanding, irrespective of model architecture.

MedHalwasa contributes not only a novel dataset but also key empirical insights into the nature of hallucinations in Arabic medical generation. These findings carry important implications for the deployment of LLMs in real-world healthcare appli-

cations, particularly in low-resource languages like Arabic. We hope our work serves as a foundation for future efforts aiming to build more trustworthy, medically grounded, and linguistically aware LLMs.

6 Limitation

The factual accuracy of generated statements was assessed using automated evaluation frameworks. While these systems offer scalable and consistent measurement, their reliability is inherently limited by their internal claim decomposition strategies, the quality of external tools used, and the scope of their underlying models' medical knowledge. As a result, they may misclassify certain nuanced medical statements, particularly in a low-resource language like Arabic. Although our evaluation provides a robust estimation of relative factuality across models, human annotation would yield more accurate assessments and remains an important direction for future validation.

Additionally, in our data generation process, we employed a single, unified prompt for all models to ensure a fair and unbiased comparison. However, LLMs may respond differently to varied phrasings, and certain prompts could better align with specific models' training distributions. This fixed prompting strategy may not reveal the upper bound of each model's capabilities, especially in prompt-sensitive scenarios. Future work could explore prompt engineering or model-specific prompting to better characterize factual performance under optimal conditions.

Finally, our study primarily emphasizes factual accuracy and does not comprehensively evaluate the linguistic fluency, coherence, or dialectal correctness of the generated Arabic text. These factors, while orthogonal to factuality, are critical for the practical adoption of LLMs in real-world Arabic healthcare applications and deserve future exploration.

References

Samir Abdaljalil, Hasan Kurban, and Erchin Serpedin. 2025. Halluverse25: Fine-grained multilingual benchmark dataset for llm hallucinations. *arXiv preprint arXiv:2503.07833*.

Angus Addlesee. 2024. Grounding llms to in-prompt instructions: reducing hallucinations caused by static pre-training knowledge. In *Joint International Conference on Computational Linguistics, Language Resources and Evaluation 2024*, pages 1–7. European Language Resources Association.

Vibhor Agarwal, Yiqiao Jin, Mohit Chandra, Munmun De Choudhury, Srijan Kumar, and Nishanth Sastry. 2024. Medhalu: Hallucinations in responses to healthcare queries by large language models. *Preprint*, arXiv:2409.19492.

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

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Chengfeng Dou, Ying Zhang, Yanyuan Chen, Zhi Jin, Wenpin Jiao, Haiyan Zhao, and Yu Huang. 2024. Detection, diagnosis, and explanation: A benchmark for chinese medial hallucination evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4784–4794.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zishan Gu, Changchang Yin, Fenglin Liu, and Ping Zhang. 2024. Medvh: Towards systematic evaluation of hallucination for large vision language models in the medical context. *arXiv preprint arXiv:2407.02730*.

Stefan Hegselmann, Shannon Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. 2024. Medical expert annotations of unsupported facts in doctor-written and llm-generated patient summaries. *PhysioNet. https://doi. org/10*, 13026.

Heyuan Huang, Alexandra DeLucia, Vijay Murari Tiyyala, and Mark Dredze. 2025a. Medscore: Factuality evaluation of free-form medical answers. *Preprint*, arXiv:2505.18452.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen,

Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Inception. 2024. Jais family model card.

Hasan Iqbal, Yuxia Wang, Minghan Wang, Georgi Nenkov Georgiev, Jiahui Geng, Iryna Gurevych, and Preslav Nakov. 2024. OpenFactCheck: A unified framework for factuality evaluation of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 219–229, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Yubin Kim, Hyewon Jeong, Shen Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo R Gameiro, and 1 others. 2025. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, pages 2025–02.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-qa: A real-world medical q&a benchmark. *arXiv preprint arXiv:2401.14493*.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix2: Bio-medical expert lmm for diverse medical modalities. *Preprint*, arXiv:2412.07769.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, and 244 others. 2024a. Openai o1 system card. *Preprint*, arXiv:2412.16720.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Peace Silly, Zineb Ibnou Cheikh, and Gracia Kaglan. 2025. Ai hallucinations in healthcare: Cross-cultural and linguistic risks of llms in low-resource languages. https://apartresearch.com. Research submission to the research sprint hosted by Apart.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *Preprint*, arXiv:2307.15343.

Jinge Wu, Yunsoo Kim, and Honghan Wu. 2024. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*.

Rojin Ziaei and Samuel Schmidgall. 2023a. Language models are susceptible to incorrect patient self-diagnosis in medical applications. *Preprint*, arXiv:2309.09362.

Rojin Ziaei and Samuel Schmidgall. 2023b. Language models are susceptible to incorrect patient self-diagnosis in medical applications. *arXiv preprint arXiv:2309.09362*.

Kaiwen Zuo and Yirui Jiang. 2024. Medhallbench: A new benchmark for assessing hallucination in medical large language models. *arXiv preprint arXiv:2412.18947*.

A Appendix

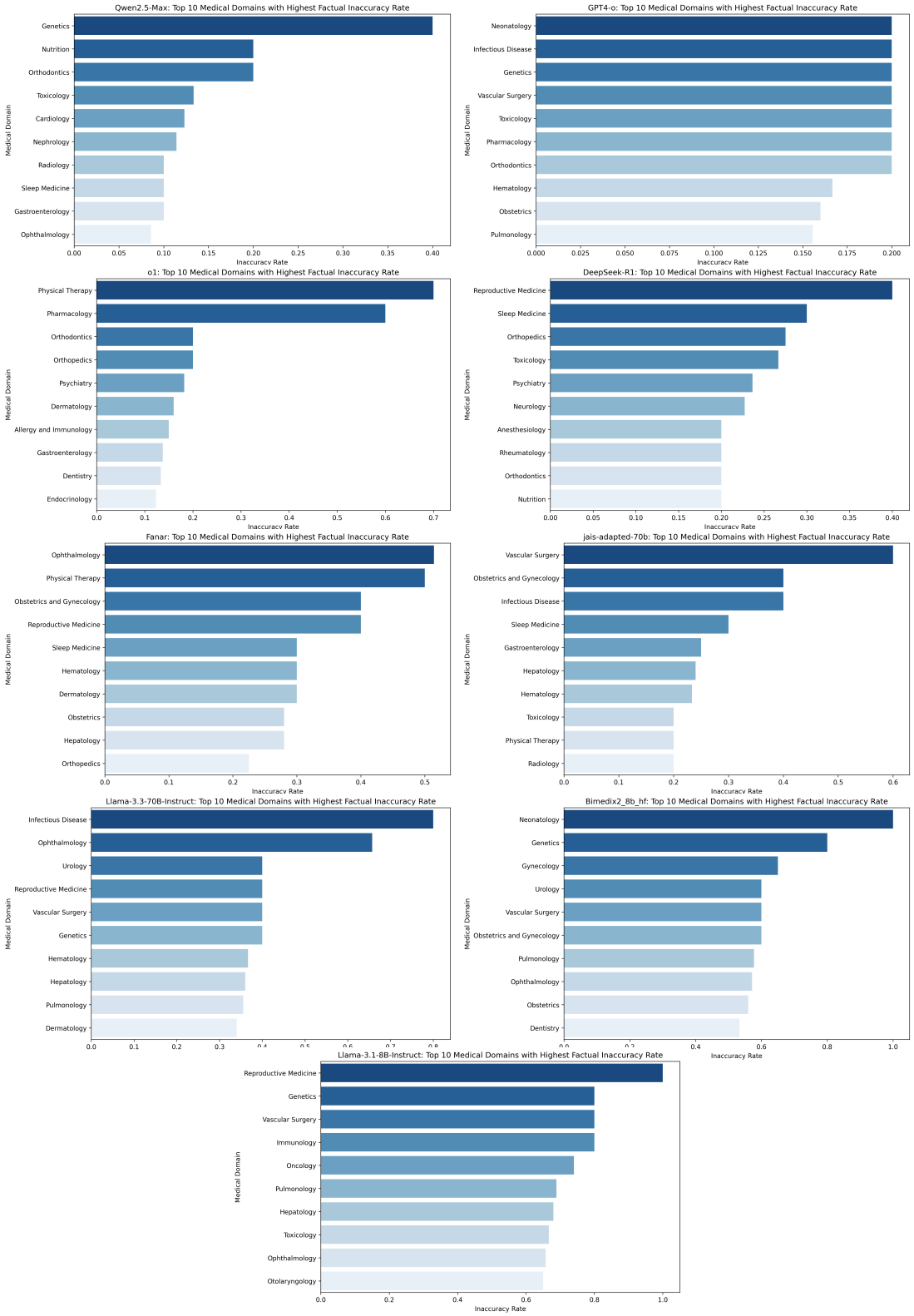

Figure 7: Top 10 medical domains with the highest inaccuracy rates in the generated content across the nine evaluated models.

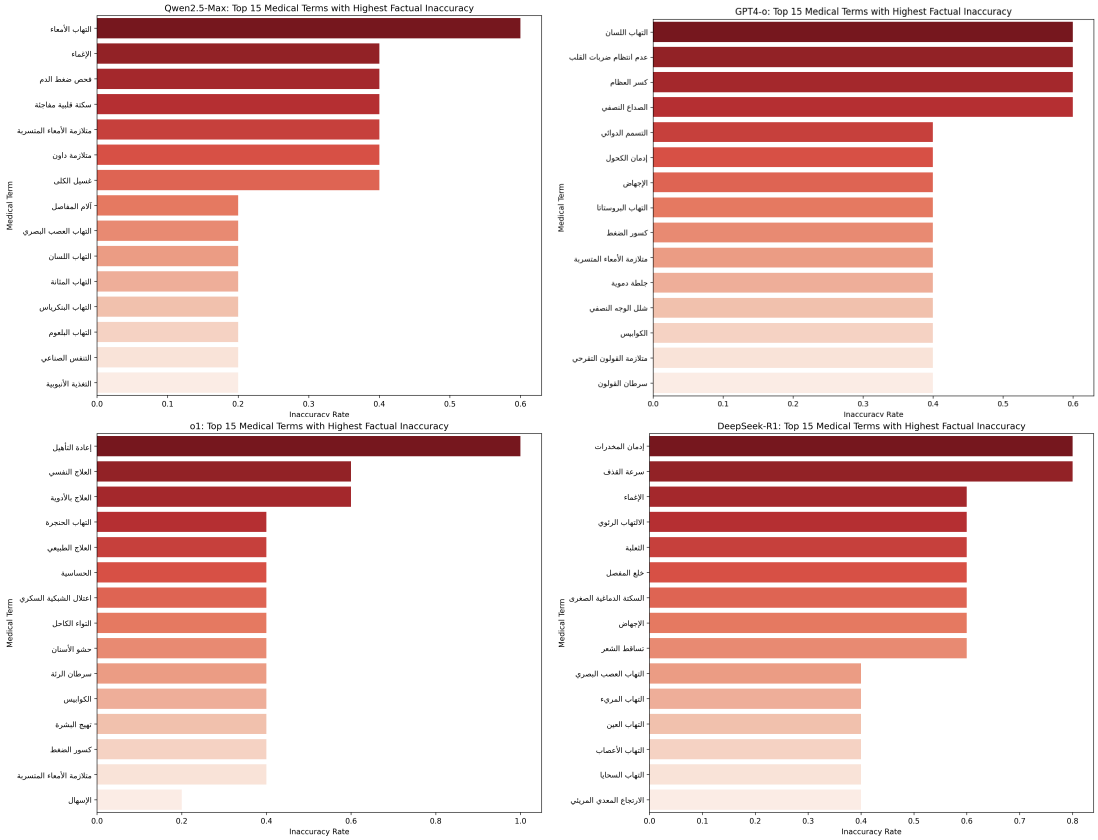

Figure 8: Top 15 medical terms with the highest inaccuracy rates in the generated content across the four large evaluated models.

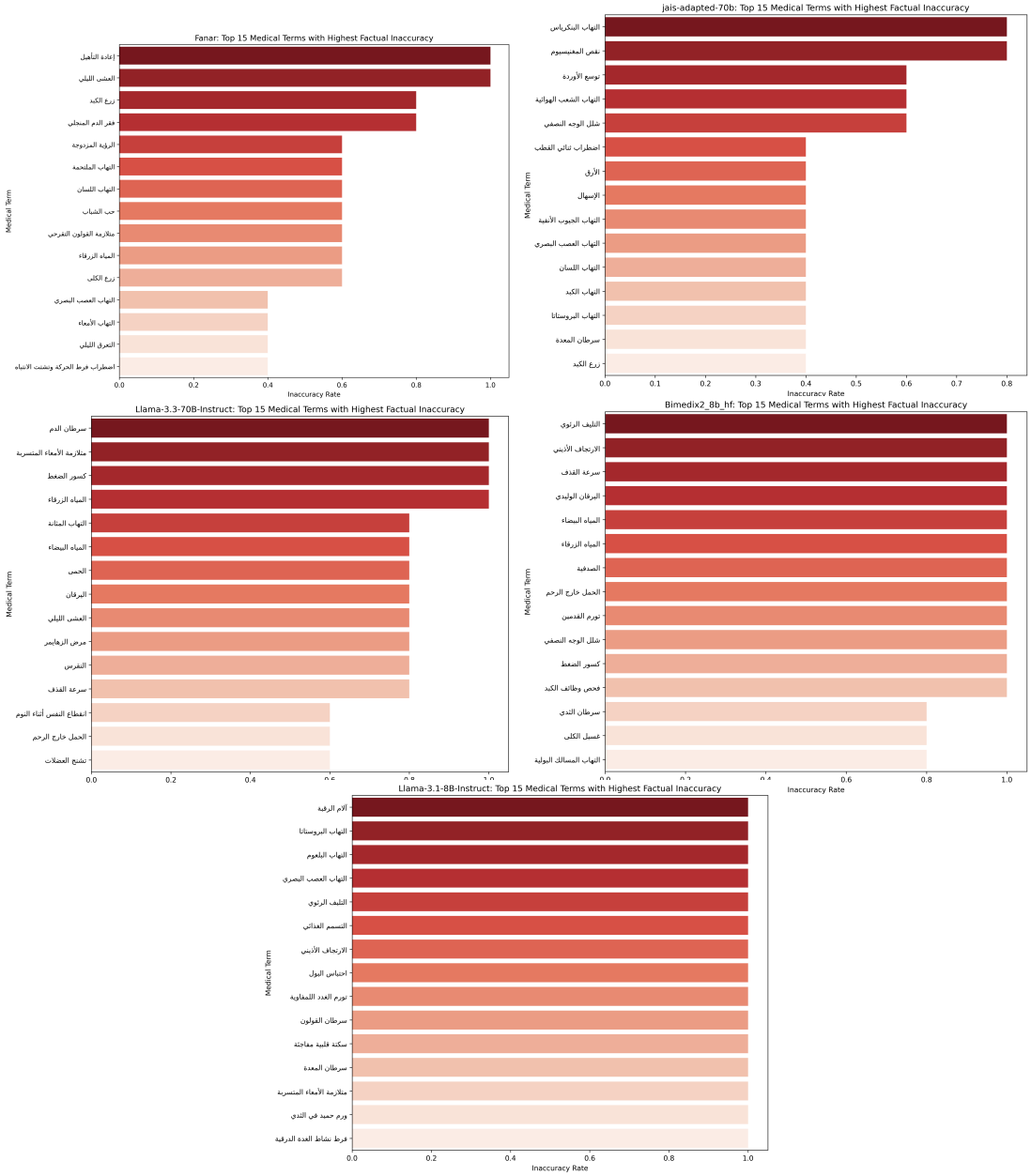

Figure 9: Top 15 medical terms with the highest inaccuracy rates in the generated content across the five medium/small evaluated models.