Leveraging Online Data to Enhance Medical Knowledge in a Small Persian Language Model

1st Mehrdad Ghassabi
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
m.ghassabi@eng.ui.ac.ir

2nd Pedram Rostami
Faculty of Computer Engineering
University of Tehran
Tehran, Iran
pedram.rostami@ut.ac.ir

3rd Hamidreza Baradaran Kashani
Faculty of Computer Engineering
University of Tehran
Isfahan, Iran
hrb.kashani@eng.ui.ac.ir

4th Amirhossein Poursina
Faculty of Medicine
Isfahan University of Medical Sciences
Isfahan, Iran
Amirhosseinpoorsina9@gmail.com

5th Zahra Kazemi
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
zhrakazemi@mehr.ui.ac.ir

6th Milad Tavakoli
Faculty of Computer Engineering
University of Isfahan
Isfahan, Iran
m.tavakoli@mehr.ui.ac.ir

Abstract—The rapid advancement of language models has demonstrated the potential of artificial intelligence in medical knowledge dissemination, yet many models struggle with specialized domains in low-resource languages like Persian. This study explores the enhancement of a small Persian language model by leveraging accessible online data, including a crawled corpus from medical magazines and a dataset of real doctor-patient Q&A pairs. We preprocess and fine-tune the model using this curated dataset to improve its medical knowledge while maintaining efficiency. Our approach focuses on domain-specific adaptation without requiring large-scale computational resources, making it feasible for localized healthcare applications. Benchmark evaluations show that the enhanced model achieves significant improvements in medical question answering accuracy and response coherence compared to its baseline. This work highlights the feasibility of leveraging open-access online data to enrich small language models in specialized fields, providing a cost-effective solution for non-English medical AI applications.

Index Terms—persian medical question answering,small language model,medical language models, data crawling

I. INTRODUCTION

The advent of the transformer architecture, as introduced in the groundbreaking paper "Attention is All You Need" [1] has catalyzed a rapid evolution in the field of natural language processing (NLP). This innovation has led to the development of increasingly sophisticated language models that leverage attention mechanisms to understand and generate human language with remarkable accuracy. As a result, the integration of artificial intelligence (AI) into various domains has surged, particularly in the medical field, where AI-driven solutions are being employed to enhance diagnostic accuracy, patient care, and administrative efficiency.

Despite the vast amount of research and development dedicated to English medical language models, such as Med-Palm [2] [3] and others, there remains a significant disparity in resources available for non-English languages, particularly Persian. To the best of our knowledge, the only existing Persian medical language model, Sina-BERT [4] , is a closed-source solution, limiting its accessibility and adaptability for further research and application. This gap underscores the urgent need for open-source resources that cater specifically to the Persian-speaking medical community.

However, this gap stems from underutilization rather than a lack of raw material. Persian-language medical forums (e.g., drhast,doctor-yab) and authoritative online magazines (e.g., hidocor, niniban) host vast amounts of expert-curated content and real-world patient-doctor interactions. These sources—if systematically crawled, cleaned, and structured—could serve as valuable resources for training a domain-specific Persian language model.

Moreover, the development of small language models is particularly crucial in the medical domain due to privacy concerns. These models can be optimized to run on local devices, ensuring that sensitive patient data remains secure and confidential, which is a paramount consideration in healthcare settings. However, the unavailability of appropriate medical corpora and datasets in Persian has hindered progress in this area, impeding the creation of robust language models that can effectively address the linguistic and cultural nuances of the Persian-speaking population.

In response to these challenges, we present a novel approach with our model, Gaokerena [1] , which fine-tunes a baseline model, aya-expanse-8b [5] , on the crawled data comprising a persian medical corpus and a medical free form farsi question answering dataset. Importantly, our model, corpus, and datasets are all open-source, promoting transparency and collaboration within the research community. This development aims to enhance access to Persian medical information and support secure, efficient interactions within the healthcare environment. By bridging the existing gaps in resources and

---

[1]Our language model is named after Gaokerena, an ancient Persian mythological tree believed to possess healing properties and grant immortality to those who consume its fruit.

leveraging advancements in NLP, our work contributes to the growing landscape of AI in medicine, particularly for Persian-speaking users.

Our contributions in this work are as follows:

- Introducing the first open-source [2] persian medical language model that achieved state of the art result in comparison to other home device runnable alternatives
- Introducing a persian medical corpus obtained by crawling different websites.
- Introducing the first persian free form medical question answering dataset obtained by crawling different websites.
- translating medical portion of MMLU bechmark which can be used to evaluate any persian medical language model

II. RELATED WORK

A. Related Works in English

Several notable projects have contributed to the development of medical language models, employing various strategies to enhance their performance and applicability in healthcare.

ChatDoctor [6], which is the most similar work to ours, represents a notable initiative focused on developing a medical language model. The team behind ChatDoctor sourced its training data from HealthcareMagic and its test data from iCliniq, compiling a total of 200,000 free-form question-answering pairs from these online platforms. They then curated the dataset by filtering answers based on their length, resulting in a final collection of 100,000 high-quality pairs. Using this dataset, they fine-tuned a LLaMA model [7] to create a system capable of delivering accurate and contextually relevant medical information. Furthermore, ChatDoctor leveraged a retrieval-augmented generation (RAG) approach, which allowed the model to access and integrate external knowledge more effectively, thereby enhancing its overall performance.

Meerkat [8] is another significant contribution in the field. This project involved extracting chains of thought from medical textbooks and fine-tuning a language model using this data, alongside other supplementary datasets. By focusing on the reasoning processes involved in medical decision-making, Meerkat aimed to create a model that not only provides information but also mimics the cognitive processes of healthcare professionals, thereby supporting more nuanced and informed interactions.

MedMobile [9] represents yet another advancement in the realm of small medical language models. This work fine-tuned the Phi-3-mini model [10] using a combination of synthetic and human-generated datasets, enabling it to achieve optimal performance tailored for mobile applications in the medical domain. By focusing on the specific requirements of mobile users, MedMobile sought to deliver a model that is both efficient and effective, ensuring accessibility to high-quality medical information on the go.

B. Related Works in Persian

As previously mentioned, there has been limited research focused on Persian medical language models, highlighting a significant gap in resources for the Persian-speaking medical community. Furthermore, existing works on Persian medical question-answering systems are entirely closed-source regarding their datasets, models, and codebases. This lack of accessible resources leaves the field largely underexplored, presenting researchers with an almost blank slate to build upon. On the other hand, all of these efforts have primarily concentrated on extractive solutions, which aim to retrieve relevant information from predefined sources, rather than employing generative approaches capable of producing context-aware responses.

Perhaps the most notable effort in this area is Sina BERT [4], which involved training a BERT model using a crawled corpus alongside Persian annotated datasets specifically developed for various tasks, including medical question answering, medical sentiment analysis, and medical question retrieval. Sina BERT is the most similar work to ours among Persian-language-focused efforts; however, it differs in that it uses a BERT model—an encoder-based language model—as its baseline. This choice limits its capability for generative AI tasks, as BERT is primarily designed for understanding and extracting information rather than generating answers.

Another notable work is the Persian Medical Question Answering System developed by H. Veisi et al. [5] Their system is structured around three main modules: question processing, document retrieval, and answer extraction. The question processing module is responsible for analyzing and refining user queries, the document retrieval module locates relevant medical documents from predefined data, and the answer extraction module identifies and extracts the most suitable answers from the retrieved content.

Similar to these two works L.Darabi [6] used models like Persian BERT [7] to retrieve relevant answers. Her approach involves finding similar questions to handle repeated queries and employs strict and lenient evaluation strategies for accurate or approximate answers. Additionally, classification methods and Named Entity Recognition (NER) are used to improve answer relevance by categorizing questions and identifying medical entities like drug and disease names.

III. BASELINE MODEL

We have chosen Aya-Expanse as our baseline model primarily due to the lack of open-source Persian medical language models, which necessitates the use of a general-purpose language model. While there are several multilingual options available, including Aya-Expanse [5], Gemma [11], Qwen [12], and PersianMind [12], we have determined that Aya-Expanse is the most suitable choice for our needs. One key reason is that the training data for the other models predominantly consists of non-Persian languages, leading to biases that may result in the generation of non-Persian text, even when specified to use only Persian. In contrast, Aya-Expanse demonstrates a robust understanding of Persian grammar and produces grammatically rich Persian text, making

it a more reliable option for our research. Furthermore, our future research plans include the capability to leverage updated parameters on Aya-Vision [12] , an extension of Aya-Expanse, which will allow us to incorporate medical images such as MRIs and CT scans as inputs, thereby enhancing our model's applicability to the medical domain.

IV. DATA

A. Corpus

As previously mentioned, there is a notable absence of publicly available Persian medical corpora specifically collected for training machine learning models. This lack of a dedicated Persian medical corpus poses a significant challenge for researchers and developers aiming to create effective models for medical applications in the Persian language. Without high-quality, domain-specific textual data necessary for training, these efforts may be hindered, ultimately impacting the development of advanced medical technologies and solutions tailored for Persian-speaking populations. To provide further insight into this issue, we have compiled a comprehensive corpus containing approximately 90 million tokens and about 100,000 articles. Additionally, I. Garcia Ferrero et al [4] . collected corpora dedicated to four languages (English, French, Spanish, and Italian), which can be compared to ours in Table I. The accompanying figure 1 illustrates the share of each magazine within our corpus, effectively highlighting the diversity of sources and underscoring the need to address gaps in available resources to foster innovation and improve health-related applications.

TABLE I
COMPARISON OF OUR CORPUS WITH CORPORA COLLECTED BY I. GARCIA FERRERO ET AL

language	no. tokens	collected by
English	1.1B	I. Garcia Ferrero et al
Spanish	950M	I. Garcia Ferrero et al
French	675M	I. Garcia Ferrero et al
Italian	143M	I. Garcia Ferrero et al
Persian	90M	us

B. Dataset

The collection of a real-world doctor-patient question-answering dataset is crucial for enhancing the capabilities of language models in the healthcare domain. Such a dataset allows models to learn valuable information derived from authentic interactions between healthcare providers and patients. By analyzing these real-world exchanges, language models can grasp the nuances of medical terminology, patient concerns, and the context surrounding healthcare inquiries. Furthermore, this dataset equips models with the ability to learn not just the factual content of responses but also the appropriate structure and tone for answering questions. This dual learning process is essential, as it enables the model to generate accurate, empathetic, and contextually relevant responses, ultimately improving patient communication and support in medical environments. In this context, Yang Liu [6] highlights several
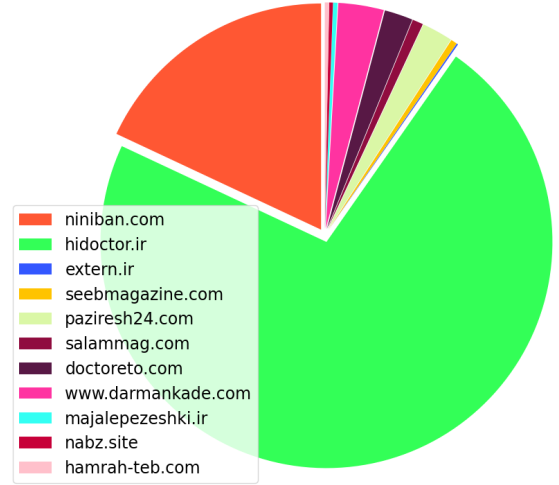

Fig. 1. Our corpus resources

real-world doctor-patient question-answering datasets in his survey, and a comparison of these datasets with ours can be found in Table II. In an era where technology increasingly aids healthcare, a robust doctor-patient dataset stands as a foundational element in training models that can effectively contribute to better healthcare delivery.

TABLE II
COMPARISON OF OUR DATASET WITH OTHERS

name	language	no. records	collected by
ChatDoctor	English	100K	Yunxiang Li et all [10]
CMtMedQA	Chinese	68K	Songhua Yang et all [10]
DISC-Med-SFT	Chinese	465K	Zhijie Bao et all [10]
HuatuoGPT-sft-data-v1	Chinese	226K	Hongbo Zhang et all [10]
Huatuo-26M	Chinese	26M	Jianquan Li et all [10]
MedDialog	Chinese & English	3.66M	Guangtao Zeng et all
Medical Meadow	English	160k	Tianyu Han et all [10]
MF3QA	Persian	20k	us

In our research, we crawled more than 180,000 question-answer pairs from Persian medical forums, employing both manual and automatic filtering methods to refine the dataset through a laborious cleaning process. This approach is similar to the work done by Yunxiang Li et al. [6] in their article on the Chat Doctor medical language model, where they extracted data from English medical forums. Notably, Yunxiang Li discarded about half of the question-answer pairs based on the length of the answers, as shorter responses are generally inadequate for training a model and can lead the model to learn to provide brief answers. However, we faced a greater challenge; Persian doctors tend to provide much shorter answers compared to their English counterparts, resulting in the necessity to discard over 80% of our question-answer records to ensure quality and relevance for our training purposes.

As you can see in the figure 2 to create our dataset, we utilized patient-doctor interactions from the drhast and niniban platforms for the training split. It is important to note that

drhast does not provide all of its doctor-patient interaction records on its site; it only offers access to the last 2,000 records. Additionally, each record is linked to 100 related records, complicating the crawling process. To address this challenge, we treated their data as a graph and performed a breadth-first search, which took about two weeks to extract 120,000 records out of a total of 200,000. For the test set, we used the doctor-yab and isovisit sites, ensuring diversity by translating the KQA question-answering dataset [10] and appending it to our test split. This comprehensive approach not only enriched our dataset but also underscored the importance of real-world doctor-patient interactions in training effective language models.

Fig. 2. MF3QA resources

V. TRAINING

A. Fine tuning

We fine-tuned the aya-expanse8b model on 60% of our corpus, leveraging LoRA [13] method to optimize performance. Specifically, we implemented a rank of 8, an alpha value of 16, and a LoRA dropout rate of 0.05, conducting the fine-tuning for a single epoch. This approach allowed us to effectively adapt the model to our specific dataset while minimizing resource usage, ensuring a balance between efficiency and performance enhancement in our training process.

B. Instruction tuning

In addition to fine-tuning the aya-expanse8b model, we also instruction-tuned the fine-tuned model on our crawled free-form Farsi question-answering (MF3QA) dataset. For this instruction tuning, we employed a LoRA method with a rank of 2, an alpha value of 2, and a LoRA dropout rate of 0.4, conducting the process for a single epoch. This targeted approach allowed us to further enhance the model's ability to understand and generate responses in the context of Farsi question-answering, optimizing its performance for our specific dataset and improving its overall effectiveness.

C. Carbon Footprint

During our recent computational experiments, we utilized an A100 PCIe 40G hardware setup on the Google Cloud Platform, specifically in the asia-east1 compute region. The hardware was operational for a total of 19 hours, resulting in a carbon footprint of 2.66 kg CO2 equivalent emissions calculated based on the formula introduced by Wu et all [14] .

VI. RESULTS

In the absence of a publicly available Persian medical language model, we opted to evaluate our model against general-purpose language models to establish a baseline for performance. This comparison allows us to assess the efficacy of our specialized model in handling medical-related queries in Persian. Importantly, all models used for comparison were selected based on their suitability for small, runnable environments on home devices, addressing privacy concerns prevalent in the medical domain. Additionally, we compared our model with a pipeline alternative in our evaluation, the architecture of which is illustrated in figure 3. By contrasting our model with both general-purpose language models and the pipeline alternative, we aim to demonstrate the advantages and specific capabilities of our small Persian medical language model in addressing the unique challenges of medical language processing within the Persian language context. [3]

Moreover, we also address the significant challenge posed by the lack of available Persian benchmark for medical language processing. To overcome this challenge, we translated the medical portion of the Massive Multitask Language Understanding (MMLU) dataset [18] into Persian and supplemented it with data from the Iranian Basic Medical Sciences Entrance Exam (IBMSEE).
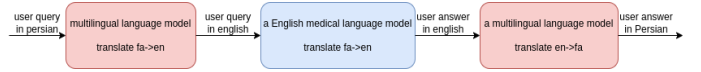

Fig. 3. architecture of a pipeline alternative

A. Comparison with general purpose language models

As you can see in Table II, our model achieved remarkable success by surpassing the passing score of the Iranian Basic Medical Sciences Entrance Exam, which stands at 36%, making it the first Persian small language model to pass this exam. Furthermore, our model demonstrated improvements on the translated MMLU dataset too, not only achieving higher average scores but also excelling across most sub-categories, thereby showcasing its effectiveness in understanding and generating medical knowledge in the Persian language. In addition to our K-QA benchmark, we also utilized GPT-4o [20] as an evaluator for free-form question answering. We provided the test set from the MF3QA dataset to both the opponent language model and our model. As illustrated in figure 4, GPT-4o predominantly preferred the responses generated by our model against all other three language models. This indicates that our model delivers high-quality responses as judged by an advanced language model.

B. Comparison with pipeline alternatives

As previously mentioned, one alternative to a Persian language model is a pipeline system. However, the major problem with pipeline systems is their speed; they have a high inference

---

[3]you can see the detailed result at https://github.com/Mehrdadghassabi/Gaokerena

|	Gao-kerena (ours)	aya-expanse8b (baseline)	Qwen2.5	PersianMind
MMLU-anatomy(fa)	**48.14**	40.74	41.48	25.18
MMLU-medical genetics(fa)	**53.0**	49.0	52.0	34.0
MMLU-college medicine(fa)	43.93	**44.51**	43.35	20.23
MMLU-clinical knowledge(fa)	**55.47**	52.07	47.92	25.28
MMLU-professional medicine(fa)	**47.05**	45.58	43.01	23.89
MMLU-college biology(fa)	**47.22**	45.14	44.85	32.63
MMLU(avg)	**49.31**	46.64	45.17	25.89
IBMSEE_Sept 2023(without time limit)	**38.69**	34.52	33.33	19.64
IBMSEE_Sept 2023(with time limit)	**38.69** (passed)	34.52 (failed)	33.33 (failed)	19.64 (failed)
inference time	low	low	low	low
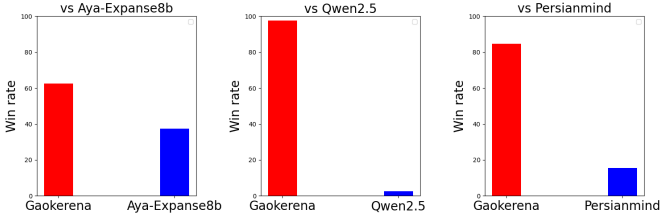
Fig. 4. Our model win rate against general purpose language models

time because they require inference from one model, the output of which is then fed into a second model, and finally, the output of the second model is processed again by the first model. This iterative process, combined with the time-consuming loading and unloading of models, significantly hampers efficiency. To address the low speed of pipeline models, one suggestion is to load all parameters—both those pertaining to the translators and the medical language model—simultaneously, thereby eliminating the delays associated with model loading and unloading. However, this approach necessitates the use of models with fewer parameters due to the limited memory available on home devices. Our experiments with models such as Medmobile [4] paired with gemma2b-it [4] as translators, and Medmobile paired with parsinlu [4] models as translators, showed disappointing results, as evidenced by the poor performance displayed in Table IV.

As you can see in Table IV, despite the impressive performance of Meerkat8b [5] paired with Aya-Expanse8b as a translator in the translated MMLU dataset and the IBMSEE exam (without a time limit), this pipeline alternative failed to

pass the IBMSEE exam under real-world conditions, which include a time limit. The pipeline system had a high inference time, taking over three minutes for each question, whereas the IBMSEE exam allocates no more than one minute per question. Another significant issue with the pipeline alternative in the medical domain is its poor performance in detecting medical terms, likely due to the fact that the translators have not been specifically developed for medical translation, as there are currently no models tailored for medical translation in the Persian language. As illustrated in Figure 5, this limitation resulted in all pipeline alternatives achieving lower win rates against our model, Gaokerena. For instance, although Meerkat8b paired with Aya-Expanse8b achieved a better score in multiple-choice question answering, in free-form question answering, GPT-4o predominantly preferred the responses generated by our model over those of the pipeline.

|	Gao-kerena (ours)	Meerkat8b + aya_expanse8b	MedMobile + gemma2b-it	MedMobile + parsinlu
MMLU-anatomy(fa)	48.14	**50.37**	14.07	25.18
MMLU-medical genetics(fa)	53.0	**62.0**	20.0	35.0
MMLU-college medicine(fa)	43.93	**53.75**	19.08	27.17
MMLU-clinical knowledge(fa)	55.47	**60.75**	27.54	31.70
MMLU-professional medicine(fa)	47.05	**57.72**	17.27	33.82
MMLU-college biology(fa)	47.22	**59.72**	18.75	31.25
MMLU(avg)	48.66	**57.57**	20.11	30.99
IBMSEE_Sept 2023(without time limit)	38.69	**56.54**	24.40	32.73
IBMSEE_Sept 2023(with time limit)	**38.69** (passed)	11.30 (failed)	24.40 (failed)	32.73 (failed)
inference time	low	very high	high	high
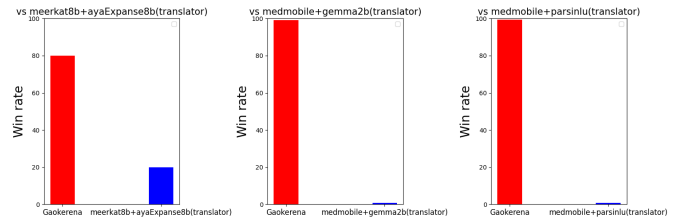
Fig. 5. Our model win rate against pipeline alternatives

FUTURE RESEARCH

As mentioned in the baseline model section, our future research plans involve leveraging our updated parameters on

Aya-Vision8b, an extension of Aya-Expanse. This will allow us to incorporate medical images, such as MRIs and CT scans, as inputs to our model, enhancing its applicability in the medical domain. By integrating textual and visual data, we aim to improve diagnostic accuracy and provide healthcare professionals with a more comprehensive decision-making tool. Future work will focus on optimizing this integration and evaluating the model's performance in real-world medical scenarios.

REFERENCES

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[2] Singhal, Karan, et al. "Toward expert-level medical question answering with large language models." Nature Medicine (2025): 1-8.

[3] Singhal, Karan, et al. "Large language models encode clinical knowledge." Nature 620.7972 (2023): 172-180.

[4] Taghizadeh, Nasrin, et al. "SINA-BERT: a pre-trained language model for analysis of medical texts in Persian." arXiv preprint arXiv:2104.07613 (2021).

[5] Dang, John, et al. "Aya expanse: Combining research breakthroughs for a new multilingual frontier." arXiv preprint arXiv:2412.04261 (2024).

[6] Li, Yunxiang, et al. "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge." Cureus 15.6 (2023).

[7] Touvron, Hugo, et al. "Llama: Open and efficient foundation language models." arXiv preprint arXiv:2302.13971 (2023).

[8] Kim, Hyunjae, et al. "Small language models learn enhanced reasoning skills from medical textbooks." arXiv preprint arXiv:2404.00376 (2024).

[9] Vishwanath, Krithik, et al. "MedMobile: A mobile-sized language model with expert-level clinical capabilities." arXiv preprint arXiv:2410.09019 (2024).

[10] Abdin, Marah, et al. "Phi-3 technical report: A highly capable language model locally on your phone." arXiv preprint arXiv:2404.14219 (2024).

[11] Team, Gemma, et al. "Gemma: Open models based on gemini research and technology." arXiv preprint arXiv:2403.08295 (2024).

[12] Rostami, Pedram, Ali Salemi, and Mohammad Javad Dousti. "Persian-mind: A cross-lingual persian-english large language model." arXiv preprint arXiv:2401.06466 (2024).

[13] Hu, Edward J., et al. "Lora: Low-rank adaptation of large language models." ICLR 1.2 (2022): 3.

[14] Lacoste, Alexandre, et al. "Quantifying the carbon emissions of machine learning." arXiv preprint arXiv:1910.09700 (2019).

[15] Team, Gemma, et al. "Gemma 2: Improving open language models at a practical size, 2024." URL https://arxiv. org/abs/2408.00118 1.3 (2024).

[16] Khashabi, Daniel, et al. "Parsinlu: a suite of language understanding challenges for persian." Transactions of the Association for Computational Linguistics 9 (2021): 1147-1162.

[17] Kashefi, Omid. "MIZAN: a large persian-english parallel corpus." arXiv preprint arXiv:1801.02107 (2018).

[18] Hendrycks, Dan, et al. "Measuring massive multitask language understanding." arXiv preprint arXiv:2009.03300 (2020).

[19] Manes, Itay, et al. "K-qa: A real-world medical q&a benchmark." arXiv preprint arXiv:2401.14493 (2024).

[20] Hurst, Aaron, et al. "Gpt-4o system card." arXiv preprint arXiv:2410.21276 (2024).