Medical and Drug Question-Answering

1st Hadis Ahmadian
Computer Engineering
Sharif University of Technology
Tehran, Iran
Ahmadiyan.h.99@gmail.com

2nd Mahsa Yazdani
Computer Engineering
Sharif University of Technology
Tehran, Iran
mah.yzdni@gmail.com

3rd MohammadReza Daviran
Computer Engineering
Sharif University of Technology
Tehran, Iran
mohammadreza.dn80@gmail.com

4th Nona Ghazizadeh
Computer Engineering
Sharif University of Technology
Tehran, Iran
nonaaghazizadeh@gmail.com

5th Sara Azarnoush
Computer Engineering
Sharif University of Technology
Tehran, Iran
azarnooshsa@gmail.com

6th Sina Abdous
Computer Engineering
Sharif University of Technology
Tehran, Iran
sina.abdoos3@gmail.com

***Abstract*—The remarkable progress of large language models (LLMs) in comprehending and responding to human instructions has captivated attention in recent years. However, their performance is predominantly optimized for English, posing limitations when applied to the medical domain. Consequently, their precision in critical tasks such as diagnoses, drug recommendations, and medical advice may fall short. Moreover, the complexities associated with training and deploying dialogue models in hospital settings have hindered the widespread adoption of LLMs. To address these challenges, this paper presents a comprehensive approach that involves data collection, model training, and comparative analysis. We gather a specialized dataset in Drug based datasets. To accommodate a broader user base, we facilitate question input in both Persian and English languages. For Persian inquiries, we leverage our dataset and employ the ChatGPT model to translate them into English, ensuring uniformity in language processing. Subsequently, we employ various techniques to identify the most suitable answer. This includes leveraging different pretrained language models. The objective is to generate accurate and contextually relevant responses to user queries. Upon obtaining the answer in English, we translate it back to Persian, ensuring effective communication with Persian-speaking users. Through our comprehensive approach encompassing data collection, model training, translation, and answer generation, we aim to overcome the limitations faced by LLMs in the medical domain. By optimizing their performance for medical tasks and facilitating multilingual communication, we strive to enhance precision and effectiveness in diagnoses, drug recommendations, and medical advice. This research paves the way for the wider adoption of LLMs in healthcare, ultimately benefiting both patients and medical professionals.**

I. Introduction and Related Works

The recent progress of large language models (LLMs) in comprehending and responding to human instructions has been remarkable. These advanced artificial intelligence systems, trained on vast amounts of text data using deep learning techniques, have demonstrated their ability to generate human-like responses. LLMs, such as OpenAI's GPT series, have the potential to revolutionize various industries, including marketing, education, and customer service, due to their capacity to process and understand large amounts of data and solve complex problems.

However, despite their impressive performance in natural language processing, LLMs like ChatGPT have not been explicitly designed for the medical domain. This limitation can lead to suboptimal precision in diagnoses, drug recommendations, and other medical advice, potentially posing risks to patients. Additionally, LLMs are typically trained in English, which restricts their ability to comprehend and respond to other languages. This language barrier hinders the accessibility of medical advice for individuals who do not speak English as their first language.

To overcome these challenges and integrate LLMs into the lives of ordinary people more effectively, it is crucial to develop medical-tailored LLMs that can be trained in multiple languages. This would not only enhance the accuracy of medical advice provided by these models but also ensure its accessibility to a wider audience.

In this project, we address these limitations by collecting data and training different models specifically for the medical domain. We receive user questions in either Persian or English and translate the Persian input into English using our dataset and ChatGPT. We then find the best answer in English and translate it back into Persian.

By adopting this approach, we aim to prevent the vulnerability of LLMs to specific prompts and improve their performance in the medical domain. Additionally, this project contributes to the development of Persian-tailored LLMs, which is a novel endeavor in the field.

While there is no similar work conducted in the Persian language, there have been related studies in other languages. Xiong et al. [9] worked on a medical question-answering project using LLMs. They created a Chinese dataset by translating an English dataset to Chinese and combining it with a Chinese dataset. They further fine-tuned and trained LLMs specifically for Chinese medical services.I Similarly, Wang et al. [10] performed fine-tuning of LLMs on Chinese medical data.

In the realm of drug recommendation systems, several studies have been conducted in the English language, utilizing different models and datasets. Begum et al. [11] developed a drug recommendation system using reviews and sentiment analysis with a recurrent neural network. Chen et al. [12] introduced large language models for biomedical natural language processing. Additionally, Fan et al. [?] explored the application of recommender systems in various domains in the era of LLMs.

These works provide insights into the application of LLMs in the medical domain and offer valuable perspectives for our research. However, our project distinguishes itself by focusing on the Persian language and addressing the specific challenges and requirements of the Persian-speaking population.

II. Datasets

In this study, we employ various datasets to inform different components of our model. A concise overview of each dataset is provided in the subsequent sections:

A. PubMed

PubMed is a free resource supporting the search and retrieval of biomedical and life sciences literature with the aim of improving health–both globally and personally. The PubMed database contains more than 35 million citations and abstracts of biomedical literature. It does not include full text journal articles; however, links to the full text are often present when available from other sources, such as the publisher's website or PubMed Central (PMC) [1]. Due to the substantial size of the PubMed dataset, exceeding 100GB, it was impractical to download the entirety of it into a Google Colab notebook. Consequently, we limited our download to the abstracts of articles published in 2023. Our initial attempt to download this data utilized the E-utilities API, but we encountered limitations on the data retrieval. As a result, we transitioned to using EDirect. We generated requests with EDirect on a Unix system and saved the retrieved data into a text file. Subsequently, we processed this text file into a structured CSV file for further analysis.

B. PubMedQA

PubMedQA is a novel biomedical question answering (QA) dataset collected from PubMed abstracts. The task of PubMedQA is to answer research questions with yes/no/maybe (e.g.: Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?) using the corresponding abstracts. PubMedQA has 1k expert-annotated, 61.2k unlabeled and 211.3k artificially generated QA instances. Each PubMedQA instance is composed of (1) a question which is either an existing research article title or derived from one, (2) a context which is the corresponding abstract without its conclusion, (3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion. PubMedQA is the first QA dataset where reasoning over biomedical research texts, especially their quantitative contents, is required to answer the questions ([2], [3]).

C. PubMed Summarization

The dataset encompasses approximately 133,000 PubMed articles accompanied by their corresponding abstracts. Given that an article's abstract inherently serves as a concise "summary," this dataset offers an invaluable resource for refining our foundational models tailored to summarization tasks. Specifically, our objective involves training models to process biomedical content, wherein the full article serves as input and the associated abstract is generated as output. However, it is imperative to acknowledge that resource constraints prompted us to work exclusively with a subset of this extensive dataset. [4]

D. Translator Dataset

This dataset aims to contribute to the advancement of medical translation technology by providing a custom dataset specifically focused on translating medical terminology related to diseases and drugs between English and Persian languages. Due to the scarcity of existing datasets in this domain, a novel approach was undertaken to generate a dataset suitable for fine-tuning translation models. This involved a combination of web scraping and synthetic data generation using the OpenAI ChatGPT API. The resulting dataset presents a valuable resource for training and evaluating translation models in the field of medical translation, particularly with a focus on disease and drug-related content.

A curated set of disease names in English and their corresponding translations in Persian were collected through web scraping of relevant medical websites. These disease names serve as the anchor points for dataset expansion.

For each disease name, the ChatGPT API was utilized to generate three synthetic medical sentences in both English and Persian. These sentences cover various aspects of diseases and drugs, including symptoms, treatments, and pharmaceutical interventions. The utilization of the ChatGPT API ensures the creation of contextually relevant and coherent content, enhancing the dataset's quality.

The resulting dataset comprises paired sentences in English and Persian, with a focus on medical content related to diseases and drugs. Each disease name in English is associated with its Persian translation, forming the basis of the dataset. For every disease name, three contextually diverse sentences are generated in both English and Persian, resulting in a comprehensive dataset for training and evaluation.

III. Implementation

A. Translator

We conducted a fine-tuning process on the translation model known as "SMaLL-100" [5] which stands for Shallow Multilingual Machine Translation Model for Low-Resource Languages. This model is designed to handle translation tasks between English and Persian languages. The purpose of this fine-tuning was to enhance the model's proficiency in translating text related to diseases. To accomplish this, I generated a specialized dataset focused on diseases.

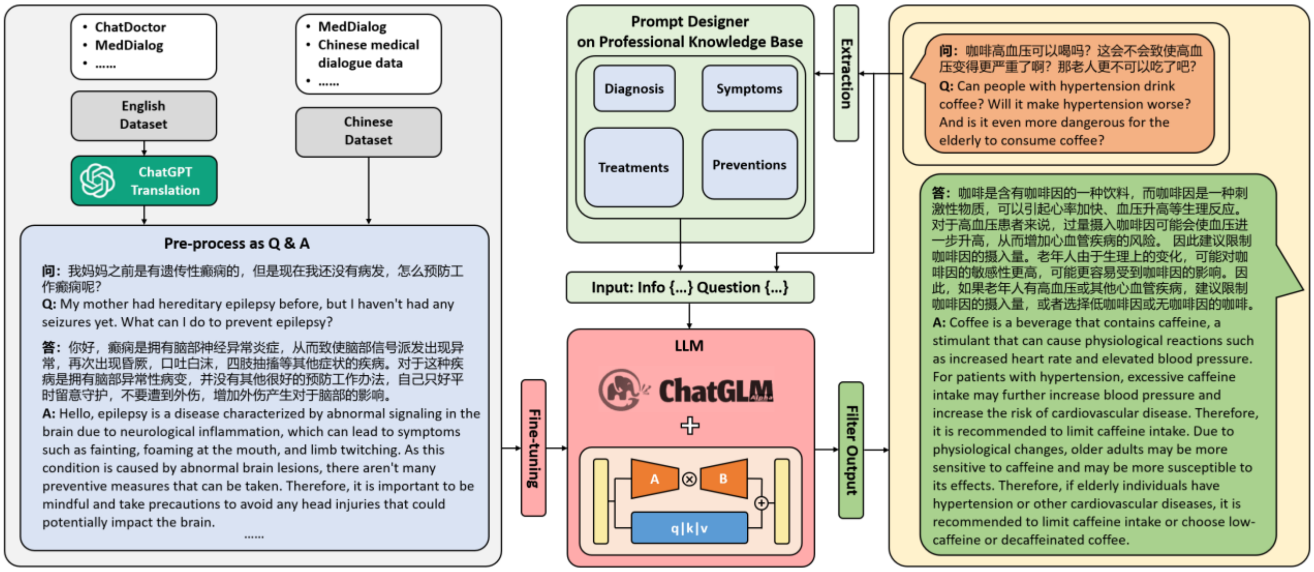

Fig. 1. DoctorGLM fine-tuning and inference pipeline

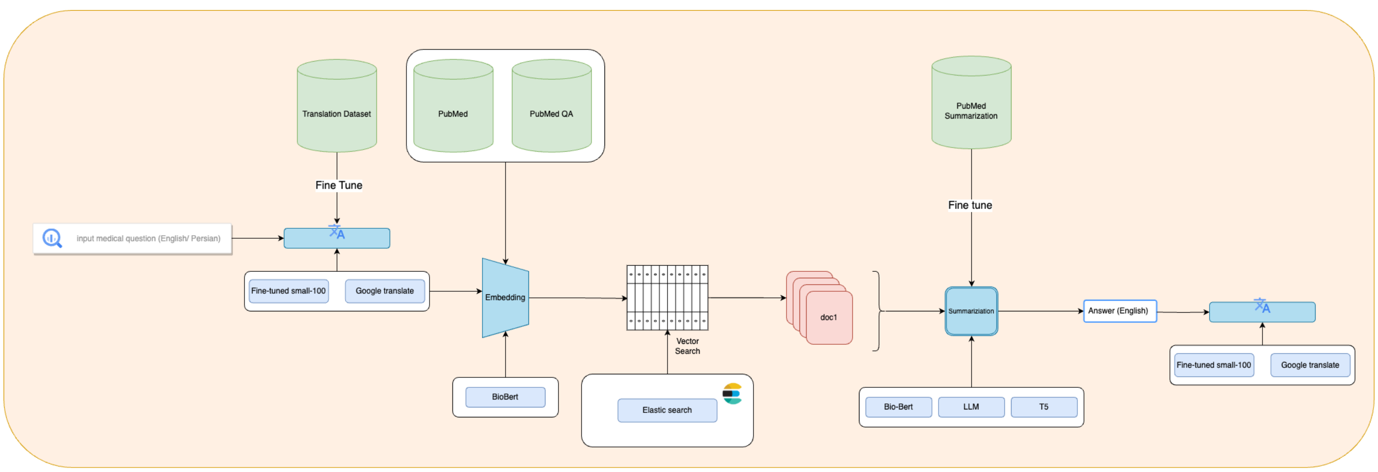

Fig. 2. Pipeline

The trained model now exhibits an improved capability to translate from English to Persian and vice versa, specifically in the context of medical conditions. This refined translation capacity is anticipated to be beneficial for tasks involving the conversion of medical information between these two languages. The model's performance was assessed based on its ability to accurately translate medical terminology and descriptions associated with diseases.

We enhance the translation capabilities between the English and Persian languages, we embarked on a fine-tuning process utilizing a specialized translation model. The focus was to facilitate accurate and contextually relevant translations within the domain of diseases and medical terminology. To this end, We conducted a dataset specifically centered around diseases, drawing upon our own resources to generate a dataset that aligns closely with the subject matter.

The model employed for this purpose is the ”SMaLL-100” [5] which stands for Shallow Multilingual Machine Translation Model designed with an emphasis on addressing the complexities of low-resource languages. This particular model architecture demonstrates a remarkable aptitude for navigating the challenges posed by languages with limited available linguistic resources.

Fine-tuning, a crucial phase in the development of machine translation models, was executed to further refine the model's language understanding and translation specialties. The dataset focused on diseases served as the foundation for this fine-tuning process. By systematically exposing the model to diverse disease-related content.

The ultimate application of this fine-tuned translation model resides in its ability to seamlessly translate textual inputs provided in either the Persian or English languages. The two-way translation functionality accommodates both the rendering of Persian text into English and the reciprocal translation of English text into Persian. In this manner, the model acts as a valuable intermediary, bridging linguistic and semantic gaps that can often emerge when dealing with specialized content such as medical discourse.

By incorporating domain-specific knowledge and training data, the model aspires to facilitate accurate, coherent, and meaningful translations, thus potentially enhancing cross-lingual communication and access to medical insights on a global scale.

B. Embedding/Vector search

After reading and processing the context from datasets, our next step involves leveraging the BioBert model. We utilize this model to embed the data, including the context, and also the translated input question. The objective is to identify the most relevant context that corresponds to the given question. Using the BioBert model, a specialized variant of BERT (Bidirectional Encoder Representations from Transformers), we generate embeddings for the prepared context data. Embeddings are numerical representations that capture the semantic meaning and contextual information of the text. These embeddings encode the relationships between words and phrases, enabling us to compare and measure their similarity.

Then we have a vector searching phase. As its name represents, we aim to search among embedding. For doing so, we first have created a database consists of our collected embeddings using Elastic Search which is a common tool in vector searching area as a result of its efficient indexing process, then we will search the query embedding among our saved ones. As mentioned earlier, the advantage of using elastic search is its indexing algorithms which are optimized for this usecase.

C. Base models for summarization

Since GPT is a large model and we didn't have resources to fine tune it on biomedical QA domain, we also tested smaller but fine-tuned language models to see if they can perform better for summarization tasks. We fine tuned BioBart and T5 large model for the task of summarization, on a dataset of 1000 pubmed articles and their abstracts (Pubmed summarization dataset).

Our fine tuned T5 model seems to have an acceptable and even comparable performance to GPT, however GPT model still outperforms fine tuned T5 and BioBart. More detail about the evaluation and comparison of T5 to GPT are provided in further sections.

D. Summarization and QA with LLMs

As we progressed through our research, we adopted a comprehensive methodology that encompassed multiple stages. Initial exploration involved the utilization of elastic search techniques to identify articles pertinent to our query. Subsequently, we harnessed the power of Language Models (LLMs) to facilitate both inference and summarization processes. However, our progress encountered a significant obstacle in the form of resource constraints. Despite our endeavors to utilize LoRA [13] for training the Falcon-7b model using the PubMedQA [2] dataset comprising 3000 PubMed articles and their corresponding abstracts, the limited scale of the dataset yielded results that failed to meet our expectations.

To address this limitation, a strategic pivot was made towards employing the GPT-3.5-turbo model API for text summarization. In this paradigm, our approach involved deploying fine-tuned BioBart and T5 models to summarize textual outputs, complemented by the GPT model's summarization capabilities. A comparative analysis was undertaken to evaluate the efficacy of these summaries. Additionally, GPT was employed for question-answering tasks, drawing from the wealth of insights garnered through summarized text. The ensuing examination of outputs unveiled a consistent trend: the utilization of GPT-based summarization consistently yielded heightened accuracy, particularly in the realm of yes/no questions sourced from the PubMedQA dataset.

In culmination, our framework presents an amalgamation of advanced techniques, namely those offered by BioBart, T5, and GPT-3.5, each contributing to the intricacies of information extraction, summarization, and question-answering. While the choice of models for summarization may vary based on preferences, we underscore the supremacy of GPT-3.5 in terms of question-answering accuracy, driven by insights culled from related articles via elastic search and cosine similarity analyses.

As we approach the final demonstration, the selection of either T5, GPT-3.5, or BioBart for summarization remains open to choice. However, it is important to emphasize that GPT remains the pivotal tool for addressing questions rooted in the content of articles identified through elastic search, establishing a cohesive connection between resource retrieval and model-driven accuracy enhancement.

IV. Experiments

A. Translation

The BLEU evaluation results suggest that the fine-tuned translation model is effective in generating accurate translations in the medical domain. The model's performance varies based on the n-gram order, with higher scores for lower n-gram orders.

The BLEU scores for various n-gram orders for translating English to Farsi are presented in the table I below

The BLEU scores for various n-gram orders for translating Farsi to English are presented in the table II below

B. Summarization

In this study, we conducted a performance evaluation of the summarization component within our pipeline using three different models: fine-tuned BioBERT, T5, and ChatGPT, on the PubMedQA dataset. To assess their effectiveness, we

TABLE I
BLEU SCORES FOR ENGLISH TO FARSI TRANSLATION

BLEU	Score
BLEU-1	0.60297
BLEU-2	0.49301
BLEU-3	0.43094
BLEU-4	0.39818

TABLE II
BLEU SCORES FOR FARSI TO ENGLISH TRANSLATION

BLEU	Score
BLEU-1	0.78158
BLEU-2	0.71473
BLEU-3	0.66086
BLEU-4	0.60995

randomly selected 10 samples from the PubMedQA dataset and provided the context portion of each sample to the models for summarization. Subsequently, we utilized ChatGPT to generate final Yes/No/Maybe answers based on the summarized text. Finally, we compared the generated final answers with the ground truth. The accuracy of the models on these 10 samples is presented in table III. Our findings indicate that BioBERT [6] exhibited the lowest performance, while ChatGPT [8] demonstrated the highest performance.

TABLE III
COMPARISON OF DIFFERENT SUMMARIZATION COMPONENTS WITH 10 SAMPLES

Model	ACC
BioBert	0.4
T5 [7]	0.6
ChatGPT	0.7

C. General Performance

To assess the overall performance of our pipeline, we conducted an evaluation using a subset of 50 questions randomly selected from the PubMedQA dataset. For each question, we proposed two article abstracts and subsequently generated summaries. Based on these summaries, we generated final Yes/No/Maybe answers. The accuracy of the pipeline, using each of the three summarization models, is presented in table IV.

TABLE IV
COMPARISON OF OVERALL PERFORMANCE WITH DIFFERENT SUMMARIZATION COMPONENTS WITH 50 SAMPLES

Model	ACC
BioBert	0.28
T5	0.39
ChatGPT	0.61

The overall comparison was conducted in English due to the limited availability of Persian resources. However, to provide illustrative examples of Persian questions and answers, figure 3 presents a selection of Persian question-answer pairs.

Question	
Medical fa_en	Google Translate
نشانه های حمله قلبی چیست	
علائم حملات قلبی شامل آنژین (درد سینه یا ناراحتی)، تنگی نفس، درد یا ناراحتی در بدن بالا (مانند دست ها، پشت، گردن، فک یا معده) است، تهوع، خستگی سر و عطسه سرما.	بر اساس حقایق پزشکی ارائه شده، علائم حمله قلبی شامل آنژین صدری (درد یا ناراحتی قفسه سینه)، ایسکمی بخش ST (که با تغییرات در بخش ST نوار قلب مشخص می شود)، افت فشار خون (فشار خون پایین) و علائم مربوط به فشار خون پایین است. مناسبت ها.
درمان سرطان ریه چیست	
بر پایه اطلاعات پزشکی ارائه شده، درمان سرطان ریه شامل شیمی درمانی، درمان تشعشعی و درمان آرامشی است. شیمی درمانی مزایای درمان پیشرفته سرطان ریه غیر کوچک را نشان داده است، به ویژه در بیماران با سرطان سلول های غیر کوچک در مراحل چهارم یا IIIB با شرکت سلول یا ن۲-ن۲ و با وضعیت عملکرد ۲ یا کمتر است. درمان سیتواستاتیک، نوعی درمان، پیش بینی مستقل از بقاء است. گزینه های درمان آرامشی برای سرطان ریه شامل درمان تشعشعی با دوز مختلف، مانند ۲۵ گی در بخش های ۵ یا ۵۰ گی در بخش های ۲۰ است. انتخاب درمان به وضعیت کلی بیمار بستگی دارد، هیچ تفاوتی در بقاء بین دو گزینه مشاهده نشده است.	بر اساس حقایق پزشکی ارائه شده، درمان سرطان ریه شامل درمان سیتواستاتیک، درمان تسکینی و رادیوتراپی است. نشان داده شده است که درمان سیتواستاتیک باعث بهبود بقا در بیماران مبتلا به کارسینوم سلول غیر کوچک در مراحل IV یا IIIB با درگیری پلور یا N2-N3 و با وضعیت عملکرد ۲ یا کمتر است. درمان تسکینی برای بیماران مبتلا به سرطان کیسه صفرا غیرقابل برداشت استفاده می شود و هدف آن بهبود علائم و کیفیت زندگی است. رادیوتراپی به عنوان درمان کمکی یا تسکین دهنده برای متاستاز ریه در کارسینوم نازوفارنکس استفاده می شود و نشان داده شده است که بقای کلی را بهبود می بخشد. انتخاب گزینه های رادیوتراپی برای سرطان پروستات به ترجیح بیمار بستگی دارد و برخی از بیماران به دلیل نگرانی در مورد سمیت، دوز کمتری از پرتودرمانی را انتخاب می کنند.
علایم سرطان ریه چیست	
علائم سرطان ریه ذکر شده در حقایق پزشکی ارائه شده، تنگی نفس، درد و دیسفاژی است.	علائم سرطان ریه ذکر شده در حقایق پزشکی ارائه شده، تنگی نفس، درد و دیسفاژی است.

Fig. 3. Persian Question-Answer samples.

V. Conclusion

In conclusion, we have presented a comprehensive pipeline for answering medical questions in both English and Persian languages. Notably, our pipeline represents a pioneering effort in addressing the medical answering task specifically in Persian. The key advantages of our model include the utilization of valid medical articles to obtain accurate answers and the generation of embeddings solely from the question, thereby minimizing security concerns related to LLM's prompting and potential biases in answer generation. For future research, we recommend leveraging our complete dataset to evaluate the model's maximum performance and fine-tuning different components using a larger corpus of Persian data.

VI. Availability

A. MedDrug-QA a web demo

As the final goal of this work was to have a question answering system for users, the need of an intractable application is clear. To address this need, we built a web application connected to our final-work-flow notebook, using anvil service. This web application is available here

References

[1] ”About PubMed.” (n.d.). National Library of Medicine. Retrieved August 14, 2023, from https://pubmed.ncbi.nlm.nih.gov/about/

[2] Hugging Face. (n.d.). PubMed QA: A Dataset for Biomedical Research Question Answering [Data set]. Retrieved August 14, 2023, from https://huggingface.co/datasets/viewer/?dataset=pubmed_qa&config=pqa_labeled

[3] ”PubMedQA.” (n.d.). Retrieved August 14, 2023, from https://pubmedqa.github.io/

[4] Hugging Face. (n.d.). PubMed Summarization Dataset: Dataset for summarization of long medical documents. Retrieved August 14, 2023, fromhttps://huggingface.co/datasets/ccdv/pubmed-summarization

[5] A. Mohammadshahi, V. Nikoulina, A. Berard, C. Brun, J. Henderson and L. Besacier, “SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages,” arXiv:2210.11621 [cs.CL], Oct. 2022.

[6] Jinhyuk Lee and others, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, Volume 36, Issue 4, February 2020, Pages 1234–1240, https://doi.org/10.1093/bioinformatics/btz682

[7] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485-5551.

[8] OpenAI. (2023). ChatGPT (Version 1.0) [Computer software]. Retrieved from https://chat.openai.com/chat

[9] Xiong, H., Wang, S., Zhu, Y., Zhao, Z., Liu, Y., Wang, Q., Shen, D. (2023). Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097.

[10] Wang, H., Liu, C., Xi, N., Qiang, Z., Zhao, S., Qin, B., Liu, T. (2023). Huatuo: Tuning llama model with chinese medical knowledge. arXiv preprint arXiv:2304.06975.
[11] Begum, S. G., Sree, P. K. (2023). Drug Recommendations Using a "Reviews and Sentiment Analysis" by a Recurrent Neural Network. Indonesian Journal of Multidisciplinary Science, 2(9), 3085-3094.
[12] Chen, Q., Du, J., Hu, Y., Keloth, V. K., Peng, X., Raja, K., ... Xu, H. (2023). Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv preprint arXiv:2305.16326.
[13] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2021). Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
[14] Fan, W., Zhao, Z., Li, J., Liu, Y., Mei, X., Wang, Y., ... Li, Q. (2023). Recommender systems in the era of large language models (llms). arXiv preprint arXiv:2307.02046.