

Who Can be Your AI Doctor?: Evaluation for Disease diagnosis on Large Language Models

Jonghyeon Kim

*Dept. of Applied Artificial Intelligence
Major in Bio Artificial Intelligence
Hanyang University
Ansan, Republic of Korea
jonghyeon@hanyang.ac.kr*

Chan-Yang Ju

*Dept. of Applied Artificial Intelligence
Major in Bio Artificial Intelligence
Hanyang University
Ansan, Republic of Korea
karunogi@hanyang.ac.kr*

Dong-Ho Lee

*Dept. of Applied Artificial Intelligence
Hanyang University
Ansan, Republic of Korea
dhlee72@hanyang.ac.kr*

Abstract—Large Language Models (LLMs) have demonstrated outstanding performance in general domain knowledge inference such as arithmetic reasoning, commonsense reasoning, and open-domain question answering. Traditional LLMs were mainly developed and serviced by big tech companies as commercial, making it nearly impossible for researchers to access parameters. However, the recent emergence of non-commercial LLMs has led to ongoing research efforts to surpass commercial LLMs in specific areas using task-specific or domain-specific LLMs. In this study, among many tasks or domains, we focus on the high-level knowledge-demanding medical domain, particularly Automatic Diagnosis System (ADS). We evaluated whether current representative LLMs can perform disease diagnosis effectively and whether they have the potential to assist doctors in clinical situations. Moreover, we also observe the overall quality of responses to whether each LLM can accurately diagnose based on understanding the patient's basic information such as age, sex, underlying diseases, and family history, and explore whether non-commercial LLMs are likely to outperform commercial LLMs in ADS.

Index Terms—Automatic diagnosis system, large language models, medical domain knowledge

I. INTRODUCTION

Large Language Models (LLMs) are powerful models trained on massive amounts of text tokens and billions of parameters. These models have the capability based on their immense power to understand the given context through In-Context Learning (ICL), without the need for additional fine-tuning. Traditional research and development of LLMs were primarily carried out by big tech companies with significant computing resources and capital due to challenges such as securing computing resources and data collection.

Prominent commercial LLMs include Google LaMDA[1] with 137 billion parameters, OpenAI GPT-3[2] with 175 billion parameters, and Google PaLM[3] with 540 billion parameters. These models have not opened the pre-trained

This work was partly supported by the Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)), the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No.NRF-2022R1F1A1073208) and Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0008691, HRD Program for Industrial Innovation).

weights and training data publicly due to the reasons like ethical or copyright issues with the data and considering the model itself as an asset. Instead, they primarily release technical reports containing performance metrics and experimental results. Additionally, these commercial LLMs were primarily focused on achieving the highest performance through large model sizes and lots of data during their research. It requires an enormous computing budget for inference thus it is virtually impossible with typical methods. Furthermore, it is nearly impossible to create task-specific models by accessing parameter levels for researchers.

Meta recently released the Large Language Model Meta AI (LLaMA)[4], an instruction-following model that surpasses the performance of existing benchmarks in most scenarios with 13 billion parameters which are about ten times fewer than GPT-3. They released the models ranging from 7 billion to 65 billion parameters in a limited open-source format. Unlike previous commercial LLMs, it could be opened to the public due to it was pre-trained using only publicly available data. There are numerous attempts in process based on it, including Stanford Alpaca[5] performed instruction-tuning via Supervised Fine-Tuning (SFT), and Vicuna[6], which is intensely trained on dialogue datasets for chatbot systems.

The general domain knowledge embedded within these LLMs has shown remarkable abilities as a general task solver in various fields including arithmetic reasoning, commonsense reasoning, and open-domain question answering[3]. Numerous studies are currently evaluating the capabilities of LLMs and exploring various ways to apply them in diverse fields. Particularly in the medical and clinical fields, ChatGPT achieved a passing score on the US Medical Licensing Exam (USMLE), and it demonstrated a high accuracy of more than 83% on average under zero-shot conditions when taking the exam with GPT-4[7]-[8]. Furthermore, the quality and empathy level of responses to patient questions were found to be higher than those of actual doctors[9]. However, exam-related questions like USMLE may not be suitable for various real clinical situations, and the capability for this exam is not perfect[8].

We focus on the field of Automatic Diagnosis System (ADS) among the diverse medical and clinical fields. ADS is a system that assists doctors by suggesting potential diseases

that might have been overlooked. It can automatically diagnose patients based on the information provided by the patient, regardless of the format of data, such as medical dialogue, medical records, X-rays, CTs, and MRIs. In this context, the system can be designed to utilize a single type of data for diagnosis or use a multi-modality approach to combine multiple data types for more accurate and diverse diagnostics. However, in this study, we limit the scope of ADS to basic patient information with a single type of data that includes symptoms, age, gender, and family history which had been completely de-identified. Such information can be used to provide more appropriate medical services to patients by checking the condition before meeting with a doctor and can also be utilized during disease diagnosis when a patient consults a doctor at the hospital. Existing ADS research has utilized external knowledge about the relation between symptoms and diseases to enhance the accuracy of diagnoses. However, knowledge-based disease diagnosis can be challenging when it comes to diseases outside of the existing knowledge base. On the other hand, LLMs have potential use as an ADS since they have been trained on more comprehensive knowledge covering various diseases and symptoms.

With these potential, evaluating the capability of disease diagnosis for LLMs is need for ADS research. Therefore, we explore how helpful the general knowledge of various LLMs is in disease diagnosis and whether they can serve as ADS to assist doctors in real clinical situations. Unlike previous studies, we measure the capabilities of disease diagnosis using actual case study data focusing not on how well the LLMs mimic doctor-like language but on how well they diagnose diseases exactly and provide information about related conditions with minimal patient information. To the best of our knowledge, there have been few efforts to evaluate the disease diagnostic capabilities of multiple commercial and publicly available LLMs using patient basic information.

II. RELATED WORKS

A. Automatic Diagnosis System (ADS)

With increasing data availability for various ADS applications and advancements in Artificial Intelligence (AI), the demand for automated disease diagnosis is rising in both research and industrial fields. Recently, learning-based ADS actively utilizes Transformer[10] architecture, which demonstrates outstanding performance in natural language processing and image processing. Various research is ongoing based on this architecture such as making the ADS data public and evaluating them; developing medical dialogue systems using patient-doctor data; and using multi-modality, including symptom-related texts and visual information for symptoms[11]-[14].

Concretely, The issue was raised that it is difficult to explain some symptoms in words and that people are not familiar with many medical terms used for symptom expression. To handle this problem, proposed a multi-modal disease diagnosis virtual assistant that diagnoses diseases using both patient symptom

images and symptom utterances[11]. In [12], COVID-19 medical dialogue data in English and Chinese was collected to develop a medical conversation system capable of providing related consultations when difficult to visit hospitals such as during a pandemic. They attempted to alleviate overfitting with a masked-token prediction due to the collected dataset was not large in quantity. In [13], pointed out that the existing reinforcement-based or non-reinforcement-based methods only focus on disease diagnosis and ignore the patient’s symptoms. For that reason, the system uses a Transformer decoder-encoder and tokenizes both symptoms and diagnoses to attend to the symptoms and diagnostic information. However, this approach diagnoses diseases based solely on symptoms, excluding basic information such as the patient’s age, sex, and family history.

Recently, there has been an increasing volume of research leveraging reinforcement learning to enhance the accuracy of disease diagnosis. In [14], built a dataset extracting symptoms from patient-doctor conversation data, and proposed a dialogue system framework for disease identification. Furthermore, some studies have used knowledge graphs between the relation of diseases and symptoms to improve disease diagnosis accuracy. However, constructing such graphs is challenging without abundant medical knowledge, and it is difficult to respond to diseases that do not exist within the graph[15]-[16].

B. Large Language Models (LLMs)

There were some problems with the initial LLMs. GPT-3 can generate text that may contain bias, hate speech, misinformation, or sensitive or confidential information since it has been pre-trained on vast amounts of data collected from the web. To relieve these problems, InstructGPT[17] and ChatGPT[18] have been improved with learning methods such as SFT and Reinforcement Learning through Human Feedback (RLHF), enabling more human-like answering and outputting answers that humans prefer. These models possess the ability to speak more human-like and provide answers that align with human preferences.

Thanks to the open-sourced release of LLaMA which is an accessible-size LLM, task-specific or domain-specific LLM research have accelerated. Particularly, research is being conducted to utilize LLMs in specific fields by instruction-tuning or SFT on task-specific or domain-specific datasets over medicine, healthcare, law, and finance[19]-[22]. This tuning process not only enables customization of LLMs specialized in specific domains according to user intent but also addresses the limitation of commercial LLMs like ChatGPT, which have been unable to learn information since September 2021, making it hard to incorporate the latest information.

In the medical field even after completing the tuning process, there are no clear metrics to measure the performance for field or task-specific LLMs. For that reason, previous research evaluated the quality of generated responses aspect of how doctor-like they are, usefulness, smoothness, safety, and other factors using the same evaluation metrics as existing text generation models. Additionally, some studies evaluated the

ability to solve standardized exams such as the US Medical Licensing Exam (USMLE)[23]-[24].

Concretely, [23] has collected 100,000 patient-doctor conversations from online medical consultation platforms and fine-tuned the smallest LLaMA model which has 7 billion parameters to answer questions about users’ natural language symptom descriptions in a manner like a real doctor’s consultation. However, this approach is limited to performing a Question-and-Answer (Q/A) role and can only provide a single answer per question, unlike a chatbot which can engage in a continuous conversation by retaining all past interactions. Furthermore, they measured the similarity between the response of the LLM and the response of the real doctor, which means that even if the diagnosis is not accurate, achieving high performance is possible if the LLM responds well in a doctor-like manner.

In [24], they employed not just the patient-doctor conversation data from [23] but also gathered many comprehensive medical data sources, including flashcards utilized by medical students and information from the Stack Exchange medical sciences platform. LLaMA models with 7 billion, 13 billion, and 30 billion parameters were fine-tuned using these datasets and measured the accuracy for the USMLE, which demands expert medical knowledge. However, these standardized test questions have limited scenarios and cases depending on the number of questions, making it challenging to comprehensively cover the diverse situations present in the clinical field.

In addition to this, some researchers are targeting to mimic the role of specialists by targeting specific sections rather than focusing on general medical knowledge. They are also automating various aspects of research, such as collection, processing, and analysis of disease-related data, through prompt-based approaches[25]. This automation aims to simplify labor-intensive tasks that were traditionally done manually and automate the entire data collection, processing, and analysis paradigm, thereby contributing to advancements in the field.

III. METHODS

A. Dataset

Due to the difficulty, sensitivity, and specificity of collecting the medical data, we have utilized publicly available medical data that has been removed critical and sensitive personal information. We evaluated the diagnostic capabilities of various LLMs using PolyMed[26] dataset, which is structured in a symptom-disease format. The dataset was collected from patient-doctor conversations from a Chinese online healthcare platform. In comparison to existing datasets for ADS, it contains a significantly wider type of diseases and symptoms, as well as a larger volume of data. It does not just contain symptom and disease information but also includes departments, age, gender, family history, the background of the symptom manifestation (e.g., intense exercise), family history, and underlying disease records. The data other than symptoms and diseases can be NULL values.

The data consists of training and testing data in highly flexible JavaScript Object Notation (JSON) format. Fig. 1. is

a sample of the training data. The test data are composed of single data, which diagnoses a single disease, multi data, which diagnoses multiple diseases, and unseen data, which diagnoses diseases not included in the training data using pre-built external knowledge. We selected the single data from the test dataset to evaluate the diagnostic performance of LLMs. Table I shows the number of training and testing data as well as the unique number of diseases and symptoms.

B. Models

We chose a total of six models for the evaluation, which encompasses well-known commercial LLMs as well as publicly accessible non-commercial LLMs for research purposes. OpenAI’s GPT-3.5 series was selected as the commercial LLMs: Text-Davinci-003, GPT-3.5-Turbo, and Text-Davinci-002. Text-Davinci-003 is a model that applied SFT and RLHF to GPT-3, GPT-3.5-Turbo is a lightweight model optimized for chat applications with SFT and RLHF applied like Text-Davinci-003, and Text-Davinci-002 is a model that applied only SFT without separate RLHF. Additionally, we evaluated the Google Bard[27], a model released relatively recently compared to the others. Bard is a lightweight and optimized version of Google’s existing LLM LaMDA, with RLHF applied for high-quality responses.

As for the opened LLMs, we selected the Alpaca model, which was fine-tuned LLaMA with various instruction sets for ICL capabilities. The Alpaca model encompasses a range of parameters, from 7 billion to 65 billion, and similar to LLaMA, larger models are known for enhancing overall performance. However, the evaluation was only carried out for the small models Alpaca-7B and Alpaca-7B LoRA[28] model due to

```
{
  "train": [
    {
      "id": 4760,
      "category": "Ophthalmology",
      "sex": "M",
      "age": 29,
      "family_history": ["Hypertension"],
      "background": ["Alcohol consumption"],
      "underlying_disease": ["Upper
        respiratory infections"],
      "diagnosis": ["Conjunctivitis"],
      "symptoms": ["Hyperemia"]
    }
  ]
}
```

Fig. 1. A sample of the PolyMed.

TABLE I
COMPOSITION OF THE POLYMED DATASET

Data Type	Number of Data	Unique Diseases	Unique Symptoms
Train	3636	57	352
Test (single)	909	57	248

constraints in computing resources. LoRA is a technique that can significantly reduce the amount of computation by projecting the query, key, and value attention matrix of Transformer self-attention into a lower rank through matrix decomposition. All evaluations requiring computation for the opened LLMs were performed on a Google Cloud Platform (GCP) environment with 4 NVIDIA 40GB A100 GPUs.

C. Metrics

To evaluate the diagnostic performance of LLMs using basic patient information, we set up the top-1, top-3, and top-5 accuracy as metrics. Considering the data was extracted from doctor-patient conversations, it may be challenging to accurately predict diseases based solely on the provided patient symptom information. Therefore, we have considered when the response and the label belong to the following conditions as a correct diagnosis:

a) Synonyms: If the predicted disease is synonymous with the label, it is considered a correct diagnosis. The list of synonyms was derived from the publicly available synonym list which is provided by Asan Medical Center, ranked 29th in the 2023 World’s Best Hospitals List[29].

b) Category Matching: PolyMed dataset includes some labels that represent extensive categories like Allergic disease, Respiratory infection, Viral infection, or combined diseases such as Laryngopharyngitis (laryngitis + pharyngitis) and Balanoposthitis (balanitis + posthitis). Since pinpointing a precise prediction for these labels can be challenging, we determined that the prediction is correct if the predicted disease is included in the diagnosis.

c) No distinction between Chronic/Acute or Allergic/Non-allergic conditions: The symptom diagnostic outcomes of the dataset were based on a doctor’s decision, either through conversations with patients. As it is difficult to determine the type or duration of a disease based solely on symptoms can be challenging, a prediction was considered correct if the same disease was predicted regardless of whether it is chronic or acute, allergic or non-allergic.

Table II refers to examples of prediction alignment and correct answers based on these various scenarios.

D. Preprocessing

To use the raw JSON data from the patient’s basic information as input for LLMs, the following preprocessing steps were taken:

TABLE II
EXAMPLE OF THE ALIGNMENT BETWEEN PREDICTIONS AND LABELS

Prediction	Label	Answer	Note
Common cold	Respiratory infection	O	Synonym
Pharyngitis	Laryngopharyngitis	O	Category matching
Roseola	Exanthem	O	Synonym
Common cold	COVID-19	X	Incorrect answer

a) Format Transformation: LLMs typically have a maximum input length, providing the data to the LLMs in its original format can be inefficient, and cannot be potentially expected to output to the desired response. Therefore, all fields in the JSON excluding those representing the patient ID, the category indicating the medical department, and the diagnosis field indicating the diagnosis result, were consolidated into a single input sequence with line breaks (\n) as separators. In cases where a field contained more than two values, they were delimited by commas. Notably, the diagnosis field’s value was extracted independently to serve as labels for evaluation purposes.

b) Instruction Tuning: Writing good instruction is essential for obtaining high-quality answers from LLMs. To measure the diagnostic performance of LLMs through patient data, we first assigned the role of a doctor to LLM. Then, we informed the LLM of the task to diagnose the top-5 diseases based on the patient’s information. And finally, we wrote emphasis points, requesting list-form answers without additional explanations and in descending order of probability. The emphasis points were written capitalized since we found that LLMs often disregarded the emphasis points when not in uppercase.

Table III represents an example of preprocessing the data in Figure 1.

IV. EVALUATION

A. Quantitative Evaluation

All responses from every LLM for the evaluation were collected as of July 15th. The responses of OpenAI’s GPT series models were collected by the official Application Programming Interface (API). As Google did not provide an official API for Bard yet, we collected the responses by manually. In Table V, the GPT-3.5-Turbo model showed the highest top-1 accuracy (33.99%), top-3 accuracy (54.56%), and top-5 accuracy (61.6%), followed by the Text-Davinci-003 model with top-1 accuracy (30.91%), top-3 accuracy (46.86%), and top-5 accuracy (54.45%). These results are currently considered too low for application in actual clinical situations.

TABLE III
INSTRUCTON AND AN EXAMPLE OF THE PREPROCESSED DATA

Instruction	Preprocessed	
	Input	Label
You are in the role of a doctor. Please diagnose the top-5 diseases following the patient’s information and LIST THEM WITHOUT ANY DESCRIPTION IN ORDER OF HIGH POSSIBILITY.\n	Sex: M\n Age: 29\n Family history: Hypertension\n Background: Alcohol consumption\n Underlying disease: Upper respiratory infections\n Symptoms: Hyperemia.	Conjunctivitis

TABLE IV
FAILED ANSWER CASES FOR EACH MODEL

Patient Information	Model	Response	Label
Sex: W Age: 31 Symptoms: Headache, Neck pain, Phlegm	GPT-3.5-Turbo	As an AI language model, I cannot provide real-time medical diagnosis or advice. ...	Laryngopharyngitis
Sex: M Age: 27 Underlying disease: Fatty liver Symptoms: Insomnia, Abdominal bloating, Abdominal pain	Text-Davinci-002	1. Fatty liver 2. Insomnia 3. Abdominal bloating 4. Abdominal pain	Gastritis
Sex: M Age: 16 Background: Masturbation Symptoms: Urinary frequency, Delayed urination	Bard	I'm a text-based AI, and that is outside of my capabilities.	Chronic prostatitis
Sex: M Age: 26 Symptoms: Foreign body sensation in the throat (globus sensation), Burping, Chest pain, Chest tightness, Neck pain	Alpaca-7B	The top 5 diseases that could be causing the symptoms are: 1. Gastroesophageal reflux disease (GERD) 2. Laryngopharyngeal reflux disease (LPR) 3. Gastroesophageal reflux disease (GERD) 4. Gastroesophageal reflux disease (GERD) 5. Gastroesophageal reflux disease (GERD)	Chronic gastritis
Sex: M Age: 2 Symptoms: Diarrhea, High fever, Rash	Alpaca-7B LoRA	Diarrhea, High fever, Rash, Influenza, Gastroenteritis.	Exanthem subitum

TABLE V
EVALUATION RESULTS FOR LARGE LANGUAGE MODELS

Model	Top-1 Accuracy (%)	Top-3 Accuracy (%)	Top-5 Accuracy (%)
Text-Davinci-003	30.91	46.86	54.45
GPT-3.5-Turbo	33.99	54.56	61.6
Text-Davinci-002	27.83	44.22	53.35
Bard	30.8	45.54	53.57
Alpaca-7B	30.03	45.21	52.91
Alpaca-7B LoRA	28.6	41.8	48.51

Even the Alpaca model with the fewest parameters, which has 7 billion parameters, showed less than a 1% difference from Bard and Text-Davinci model without fine-tuning. Better performance can be measured if the evaluation is performed in a model with more parameters. Unlike commercial LLMs where direct parameter access is virtually impossible, pre-trained weights of Alpaca are available to researchers that enable fine-tuning. Therefore, the potential to improve diagnostic performance is limitless.

Moreover, we attempted full parameter tuning on the Alpaca-7B model using training data of the PolyMed, rather than simply evaluating diagnostic ability. However, training was difficult due to the small number of training data (3,636)

relative to the number of parameters (7 billion), and easily overfitted on the training data. We also attempted to fine-tune a fraction of approximately 0.12% (8 million) of the total learnable parameters using LoRA. However, we observed only a marginal 1% enhancement in diagnostic accuracy upon proceeding with the evaluation.

B. Qualitative Evaluation

Finally, we observed the overall trend of how well the LLMs understand patient information and instructions and can provide responses well. Table IV shows the failure cases providing diagnoses for each LLM except Text-Davinci-003. For GPT-3.5-Turbo and Bard, they made some outputs that could not perform the doctor role. Furthermore, some responses for GPT-3.5-turbo and most of the responses for Bard did not understand the part of the emphasis points, "WITHOUT ANY DESCRIPTION" of the instruction, provided additional explanations.

Text-Davinci-003 consistently outputs the five diseases for all inputs while understanding instructions and patient symptoms. However, many predictions were not included in the correct answers. Some responses from Alpaca-7B failed to predict the five diseases or contained the same disease multiple times in the responses. Text-Davinci-002 and Alpaca-7B LoRA output merely repeated the patient's condition and failed to diagnose a disease based on the given information.

In summary, GPT-3.5-Turbo was the model with the highest diagnosis accuracy nevertheless could not provide consistent answers. The Text-Davinci-003 provided answers in a consis-

tent format but had slightly lower prediction accuracy. The remaining models exhibited slightly less adeptness in processing inputs that integrated instructions with patient information, in contrast to the performance of the two models.

V. CONCLUSION

In this paper, we evaluated the disease diagnosis capabilities and observed the output quality of various LLMs using the patient basic information dataset PolyMed. The model with the highest diagnostic ability among LLMs was GPT-3.5-Turbo, showing an accuracy of 33.99% for the top-1, 54.56% for the top-3, and 61.6% for the top-5. This result indicates that the general knowledge of LLMs can help to understand the symptoms and basic information about patients to some extent and make a diagnosis but seem hard to assist doctors in actual clinical situations. However, as commercial LLMs are continuously updated to improve answer quality and expand capacity, and with new commercial LLMs being released day by day, we believe that continuous evaluation is needed. The qualitative evaluation also revealed differences in the comprehension of user instructions between public and commercial LLMs, which may be challenging for researchers to address when utilizing open-source LLMs. Lastly, given the difference in diagnostic accuracy between commercial and non-commercial LLMs, we expect that open-source LLMs with more parameters can be outperformed in real-world clinical situations by combining existing general knowledge and task-specific or domain-specific knowledge. We plan to explore more diverse medical-targeted data and collect more patient data to study LLM for ADS at a clinical level in the future.

REFERENCES

- [1] R. Thoppilan et al., “LaMDA: Language Models for Dialog Applications,” 2022, arXiv:2201.08239.
- [2] T. B. Brown et al., “Language Models are Few-Shot Learners,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901.
- [3] A. Chowdhery et al., “PaLM: Scaling Language Modeling with Pathways,” 2022, arXiv:2204.02311.
- [4] H. Touvron et al., “LLaMA: Open and Efficient Foundation Language Models,” 2023, arXiv:2302.13971
- [5] R. Taori et al., “Alpaca: A strong, replicable instruction-following model.” Mar. 2023. [Online]. Available: <https://crfm.stanford.edu/2023/03/13/alpaca.html>
- [6] W.-L. Chiang et al., “Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.” Mar. 2023. [Online]. Available: <https://lmsys.org/blog/2023-03-30-vicuna/>
- [7] T. H. Kung et al., “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models,” *PLOS Digital Health*, vol. 2, no. 2, p. e0000198, Feb. 2023, doi: 10.1371/journal.pdig.0000198
- [8] H. Nori, N. S. P. King, S. M. McKinney, D. Carignan, and E. Horvitz, “Capabilities of GPT-4 on medical challenge problems,” 2023, arXiv:2303.13375.
- [9] J. W. Ayers et al., “Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum,” *JAMA Internal Medicine*, vol. 183, no. 6, p. 589, Jun. 2023, doi: 10.1001/jamainternmed.2023.1838.
- [10] A. Vaswani “Attention is all you need,” *Advances in neural information processing systems*, 30.
- [11] A. Tiwari, M. Manthena, S. Saha, P. Bhattacharyya, M. Dhar, and S. Tiwari, “Dr. can see: towards a multi-modal disease diagnosis virtual assistant,” *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, Oct. 2022, doi: 10.1145/3511808.3557296.
- [12] M. Zhou et al., “On the generation of medical dialogs for COVID-19,” In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conf. on Natural Language Processing (Volume 2: Short Papers)*.
- [13] W. Chen, C. Zhong, J. Peng, and Z. Wei, “DxFormer: a decoupled automatic diagnostic system based on decoder–encoder transformer with dense symptom representations,” *Bioinformatics*, vol. 39, no. 1, Nov. 2022, doi: 10.1093/bioinformatics/btac744.
- [14] Z. Wei et al., “Task-oriented dialogue system for automatic diagnosis,” In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 201-207).
- [15] L. Xu, Q. Zhou, K. Gong, X. Liang, J. Tang, and L. Li, “End-to-End Knowledge-Routed Relational Dialogue System for Automatic Diagnosis,” In *Proc. of the AAAI conference on artificial intelligence*, vol. 33, no. 01, pp. 7346–7353, Jul. 2019, doi: 10.1609/aaai.v33i01.33017346.
- [16] S. Lin et al., “Graph-Evolving Meta-Learning for Low-Resource Medical Dialogue Generation,” In *Proc. of the AAAI conference on artificial intelligence*, vol. 35, no. 15, pp. 13362–13370, May 2021, doi: 10.1609/aaai.v35i15.17577.
- [17] L. Ouyang et al., “Training language models to follow instructions with human feedback,” 2022, arXiv:2203.02155.
- [18] OpenAI, <https://chat.openai.com/>, Retrieved in 2023, online.
- [19] Z. Liu et al., “PharmacyGPT: The AI Pharmacist,” 2023, arXiv:2307.10432
- [20] Z. Liu et al., “Radiology-GPT: A Large Language Model for Radiology,” 2023, arXiv: 2306.08666
- [21] Q. Huang et al., “Lawyer LLaMA Technical Report,” 2023, arXiv: 2305.15062
- [22] S. Wu et al., “Bloomberggpt: A large language model for finance,” 2023, arXiv:2303.17564.
- [23] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang., “ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge,” *Cureus*, 15(6).
- [24] T. Han, et al., “MedAlpaca—An Open-source Collection of Medical Conversational AI Models and Training Data,” 2023, arXiv: 2304.08247
- [25] H. Dai et al., “AD-AutoGPT: An Autonomous GPT for Alzheimer’s Disease Infodemiology,” arXiv: 2306.10095
- [26] PolyMed: A Medical Dataset Addressing Disease Imbalance for Robust Automatic Diagnosis Systems, Zenodo, Apr. 2023, doi:10.5281/zenodo.7866103.
- [27] J. Manyika, “An overview of Bard: an early experiment with generative AI.” 2023. [Online]. Available: <https://ai.google/static/documents/google-about-bard.pdf>
- [28] E. J. Hu et al., “Lora: Low-rank adaptation of large language models,” 2021, arXiv: 2106.09685
- [29] “World’s Best Hospitals 2023,” Statista. <https://r.statista.com/en/healthcare/best-hospitals-world-2023/ranking/> (accessed July. 26, 2023).