A LLMs-assisted Framework for Parkinson's Disease Assessment Based on PPMI Dataset

Zhenyu Gao[1], Qin Ni[2], Wen Liu[1], and Lei Zhang[1], *Member, IEEE*
[1]College of Information Science and Technology, Donghua University, Shanghai, China
[2]The Key Laboratory of Multilingual Education with AI, Shanghai International Studies University, Shanghai, China
lei.zhang@dhu.edu.cn

2024 7th International Conference on Algorithms, Computing and Artificial Intelligence (ACAI) | 979-8-3315-2931-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ACAI63924.2024.10899636

***Abstract*—The assessment of diseases such as Parkinson's disease, which are chronic neurodegenerative conditions, is an extremely complex and time-consuming process. Self-assessment scales or interviews in the outpatient often lead to the loss of key information. Large Language Models (LLMs) show abilities in capturing subtle linguistic differences and handling long texts, enabling the extraction of a significant amount of key information while ensuring data integrity and user privacy. We focus attention on people with Parkinson's disease (PwP). In response to these issues, we propose a framework in this paper. Firstly, we use large language model (LLM) with chain-of-thought (CoT) prompt to rephrase patient self-report texts from real-outpatient structured data as a supervised fine-tuning (SFT) corpus. Secondly, we fine-tune bidirectional encoder representation from transformers (BERT) with Low-Rank Adaptation (LoRA) to enable it to understand and extract the semanteme of self-report, thus predicting each MDS-UPDRS item. Finally, we designed experiments on a large amount of test data to evaluate the effectiveness of the framework. The results indicate that the accuracy on this task has been improved to 95.36%, which is a 6.7% increase compared to the best-performing model.**

***Index Terms*—Large language models, BERT, Parkinson's disease, Healthcare, Supervised Fine-Tuning**

I. Introduction

Since the advent of ChatGPT, the potential of its application across various tasks has presented a possible path towards artificial general intelligence (AGI). The deployment of large language models in the healthcare and medical is one of the current research attention. Healthcare and medical and concerns everyone's life, and research in this field, including Med-PaLM, BioMedLM, DoctorGLM, Med-Gemini, and other LLMs [1]–[4], have demonstrated their outstanding capabilities in the healthcare. These applications span a wide range of tasks, including medical question answering, medical advice generation, assistance in clinical diagnosis, and processing of massive amounts of medical data [5].

Parkinson's disease (PD) is the second most common neurodegenerative disorder, following Alzheimer's disease. The prevalence of Parkinson's disease may be growing faster than any other neurological disorder globally [6]. Over the past two decades, the prevalence of Parkinson's disease has significantly increased compared to the period from 1980 to 2003, with an incidence rate reaching 9.34 cases per 1000 people aged 60 and older [7]. Additionally, there is a long prodromal period before clinical manifestation of Parkinson's disease [8]. Therefore, integrating LLMs technologies into the automated diagnosis and treatment process for PD is an urgent research question.

As a result, we surveyed some neurologists, and they raised several problems in the real-world diagnosis and treatment of Parkinson's disease: 1) Cumbersome scale assessments: The diagnosis of PwP always involves a lot of scales, such as MDS-UPDRS, etc. These scales contain a number of questions, requiring much time and energy from both patients and doctors to complete, which is not only time-consuming but also cause fatigue. 2) Symptom fluctuation: Symptoms of PwP may fluctuate throughout the day or on different days, which can lead to unstable results in scale assessments and prevent tracking in time. 3) Follow-up and monitoring: Parkinson's disease is a chronic condition that requires long-term follow-up and monitoring. For doctors, tracking the progression of the disease and the effectiveness of treatment is an complex task.

In response to the said problems, in the paper, we took advantage of LLMs to conduct research based on real MDS-UPDRS data, and developed a framework that aligns medical scales using patient self-reported texts. The MDS-UPDRS has been the most widely used scale in various settings of clinical and research practices [9]. Our work is primarily divided into two steps, data generation and results prediction, as shown in Fig. 1. First, we need to obtain high-quality unstructured data, which in this case is text data, and ensure it is aligned with the real participant structured data we obtained from Parkinson's Progression Markers Initiative (PPMI). We utilized the few-shot learning ability of LLMs, guiding the model step by step to generate corpora that conform to the patient self-report format through CoT and few-shot prompting, and also provided some interpretability during the generation process. For model training and prediction, we used a model named ClinicalBERT [10] as the base model and supervised fine-tuned it using the corpora obtained from the data generation step and the LoRA framework. Finally, we transferred the task to other models, including BERT, GPT3.5, ChatGLM3-6B, and ChatGLM4 [11]–[13], and compared their performance with the framework mentioned in this paper to validate the feasibility of our framework.

The key contributions of this work are as follows:

- Utilizing the few-shot learning ability of LLMs and the chain-of-thought prompt engineering, we transformed structured data into self-reported form corpora. Based on

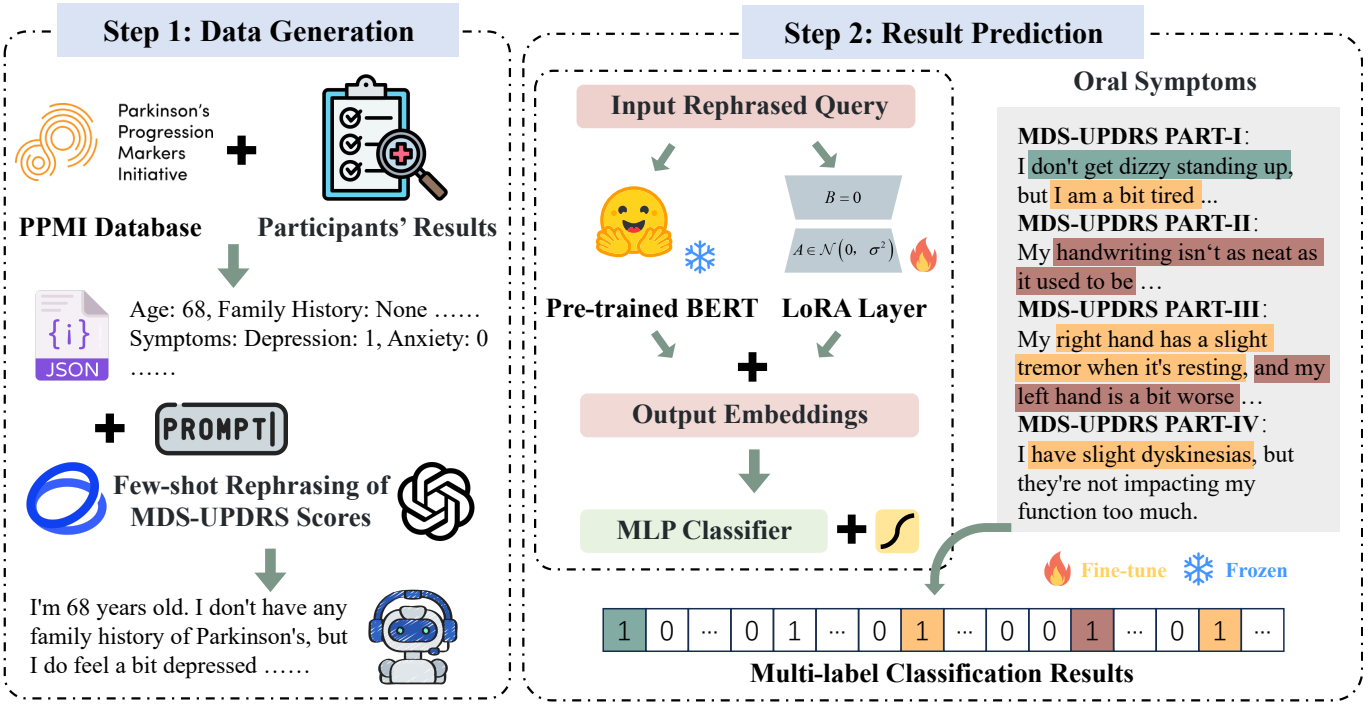

Fig. 1: The overall workflow of our framework: from outpatient data extraction to MDS-UPDRS scores prediction.

real structured table data, we formed a pipeline for generating supervised fine-tuning instructions in a human-in-the-loop manner.

- Using a pre-trained BERT as the base model, we conducted supervised fine-tuning through LoRA. We obtained a fine-tuned model capable of handling unstructured input tasks and outputting structured Parkinson's scale scores, which improved accuracy for the task compared to other models.

II. Related Work

LLMs have reshaped the application of AI in the healthcare and medical. In recent years, a variety of tools and applications related to large models have emerged. Rasmy et al. [14] proposed Med-BERT, a BERT model pre-trained on a large number of electronic health records (EHR), which introduces the successful experience of BERT in the natural language processing NLP to clinical tasks, providing a powerful pre-trained model for healthcare.

In the realm of healthcare question-answering systems, the LLM named Med-PaLM2 [15] achieved a score as high as 86.5% on the USMLE dataset, which is more than 19% higher than Med-PaLM [1] and sets a new technical standard.

Chatdoctor [16] is a medical assistant based on LLaMA [17] that has been fine-tuned using a large real-world dataset of doctor-patient conversations and integrated with a self-directed information retrieval mechanism, possessing smooth conversational abilities. Similarly, DoctorGLM [3] uses ChatGLM6B as the base model and has been trained on a Chinese medical dialogue dataset, achieving better results in Chinese scenarios.

We then focused on the application of LLM in the Parkinson's disease. Rahman et al. [18] established a user-centered teleneurology platform and assessed the possibility of using artificial intelligence technology to screen for PwP.

Among the aforementioned works, BERT is difficult for medical professionals to use directly and is more of a tool for data scientists. The generated text from healthcare LLMs can create a "trust" problem. Additionally, in the field of Parkinson's disease applications, traditional machine learning techniques like XGBoost and SVM are still used for basic identification, while LLMs mainly focus on providing chat services after obtaining the results. Our framework uses the powerful semantic classification capability of BERT as a classifier and does not provide medical advice. Moreover, it can use unstructured input instead of structured feature input, offering a more convenient user interface.

III. Materials and Methods

A. Chain-of-Thought for Data Generation

In this phase, we used real participant data from the PPMI database [19], including comprehensive MDS-UPDRS data along with demographic information such as age, gender, family history, etc. These data were compiled into JSON files. Utilizing the comprehension and the reasoning abilities of large language models (LLMs), we rephrased raw data into conversational language by simulating a outpatient scenarios. Fig. 2 provides a visual representation of the associated workflow.

Due to the extensive number of items in the MDS-UPDRS, more than 50 items, ensuring that the rephrased results

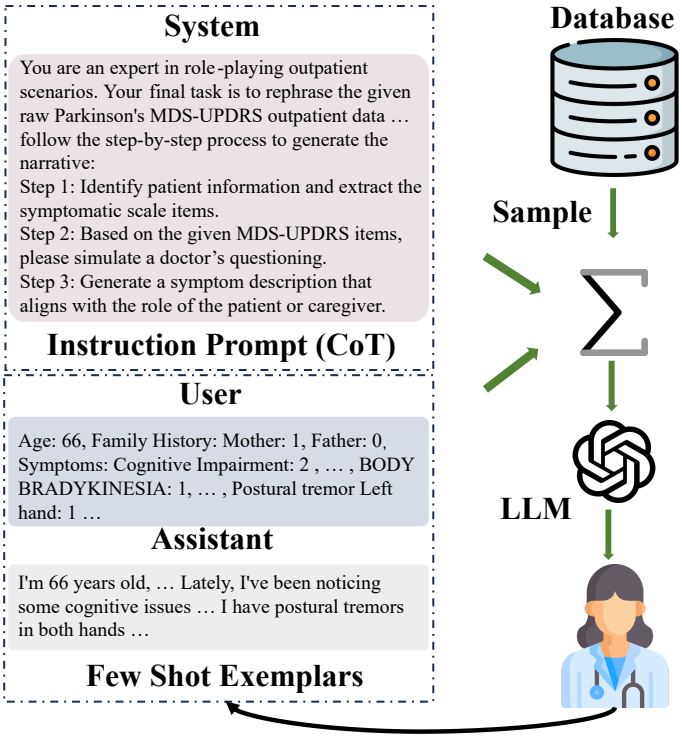

Fig. 2: The workflow of Chain-of-Thought for data generation in our framework.

accurately correspond to the true information provided by the subjects is challenging. The standard prompting method tends to hallucination, and the generation process complicates adjustments by prompt engineers. To address this, we opted for a Chain-of-Thought (CoT) approach with few-shot Exemplars as our prompting strategy. CoT has been demonstrated to bring out emergent abilities in LLMs [20]. Moreover, its step-by-step generation process enhances the controllability and interpretability of the transcription process, revealing the thought process of the LLMs and avoiding a completely non-transparent "black-box" process.

Under the instruction of system prompts, the LLM was tasked with identifying key items. Subsequently, in the model simulation of a clinic, a doctor uses a scale to conduct an interview and score a patient. The LLM generated a symptom description that aligns with the role of the patient. Through CoT, we have established a human-in-the-loop prompting engineering framework in which researchers actively participate in the generation process, recheck the generated results, provide feedback on prompt modifications, and ultimately generate more diverse and accurate results.

B. Supervised Fine-Tuning by LoRA

To better use the data generated in section 3.1 for scoring tasks, we selected a pre-trained model from the medical field named ClinicalBERT as the base model in [10]. This is a pre-trained model initialized based on BERT and trained on a large corpus of various diseases totaling 1.2B words. This pre-trained model has better generalization performance in the medical field compared to the basic BERT. Fine-tuning on this basis can adapt to our task more quickly, making it a more low-carbon, cost-effective, and faster approach.

We used the high-quality data generated in section 3.1 as the training corpus, which has good consistency with real outpatient data. Additionally, we encoded the scores of items for each sample as labels and chose the LoRA method for supervised fine-tuning. Furthermore, the final classification result of each sample corresponds to multiple labels. For example, in the sample statement "I often feel sleepy during the day, and my anxiety has been very serious recently." the output classification results should at least include MDS-UPDRS 1.8 and MDS-UPDRS 1.4. Therefore, we define this classification problem as a multi-label classification problem and set up a classification layer that fits this task. We freezed all parameters of the pre-trained model and placed the LoRA layers on the Q, V matrices, significantly reducing the number of trainable parameters using LoRA.

C. Interact Mode

In the interaction process, on one end of the interaction is the patient or caregiver, who relays the Parkinson's patients health condition and symptoms to the system in home. This process is similar to a patient visiting a clinic and consulting with a professional physician. After receiving the description, the system returns results aligned with the MDS-UPDRS scores and records the time and specifics of this assessment. Subsequently, the patient's attending physician can review the results of each home assessment, checking the efficacy of treatment or the progression of the condition based on score changes. In short, this framework allows patients to reduce the time spent on each scale by shifting from manually recorded scales to those derived from verbal narration. Particularly for Parkinson's patients, who may have lost some motor abilities or suffer from abnormal emotional fluctuations, narration is a more convenient and natural method, lightening their burden. Additionally, the MDS-UPDRS scale has many items, and it is easy for both doctors and patients to experience survey fatigue, which can affect the scoring results [21].

IV. Experiments and Results

A. Data Analysis

Our assessment data is sourced from PPMI, a real-world database specifically for Parkinson's disease and Parkinsonian syndromes. The database has collected within-participant data from over 4,000 participants across approximately 50 sites worldwide, with a gender distribution of 54.5% male and 45.5% female. For our research objectives, we selected the MDS-UPDRS from the Motor Assessments. Through LLM, we ultimately obtained a total of 60,342 labeled data for training and testing from visits of participants in the database. During the training process, the training and validation sets were divided in a ratio of 4:1.

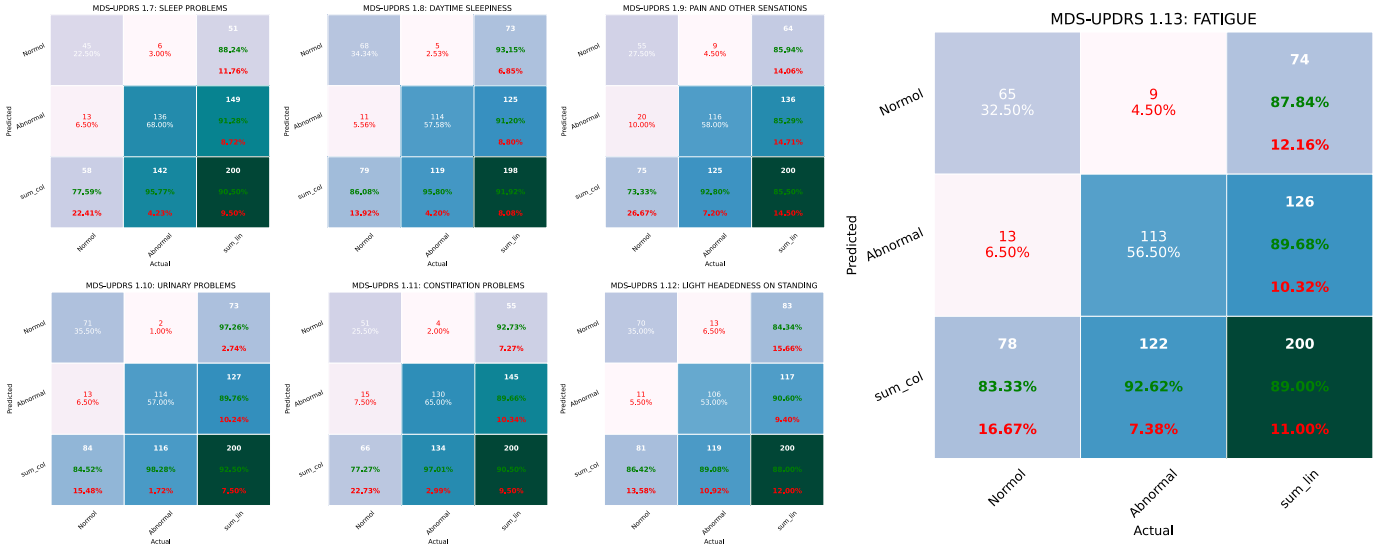

Fig. 3: The confusion matrix for each symptom in MDS-UPDRS Part I. Each subplot is derived from our sample of 200 cases, with the titles indicating each item, including sleep problems, daytime sleepiness, pain and other sensation, urinary problems, constipation problems, light headedness on standing, and fatigue.

TABLE I: Experimental parameter settings.

Parameters	
Learning rate	1×10^{-4}
Epoch	100
Batch_size	64
LoRA_rank	4
LoRA_alpha	2
Early stop patience	10
Early stop monitor	val_loss
Lr_scheduler_type	Cosine Annealing
Trainable parameters	7.4M
Non-trainable parameters	135M

B. Training Settings

Our training process was run on a server with Ubuntu 22.04, Python 3.10, and a single RTX3090.

During the training process, we did not select full parameter fine-tuning. Instead, we chose LoRA fine-tuning. Compared to full parameter fine-tuning, LoRA better preserves the prior knowledge of the pre-trained model and offers higher computational efficiency. By freezing the parameters of the pre-trained model, we kept the number of trainable parameters in the model to 7.4M. We set the learning rate to 1×10^{-4} and used Cosine Annealing as the learning rate decay strategy, with LoRA_rank set to 4 and LoRA_alpha set to 2. All the training parameters are as shown in Table I.

C. Model Evaluation Results

With the trained model obtained, we selected a variety of evaluation methods. Since our problem is a multi-label classification, where one sample corresponds to multiple labels, the common methods for multi-classification models are not applicable. Therefore, we referred to the evaluation methods for multi-label classification as described in [22], using Label-based metrics to calculate TP, TN, FP, FN. The j-th class labely_j can be calculated in (1).

$$\begin{aligned} TP_j &= |\{x_i \mid y_j \in Y_i \wedge y_j \in h(x_i), 1 \le i \le p\}|; \\ FP_j &= |\{x_i \mid y_j \notin Y_i \wedge y_j \in h(x_i), 1 \le i \le p\}|; \\ TN_j &= |\{x_i \mid y_j \notin Y_i \wedge y_j \notin h(x_i), 1 \le i \le p\}|; \\ FN_j &= |\{x_i \mid y_j \in Y_i \wedge y_j \notin h(x_i), 1 \le i \le p\}|. \end{aligned} \tag{1}$$

Where p is the total number of instances involved in the evaluation, $h(\cdot)$ is the classifier.

Based on the above four quantities, we can compute most of the binary classification metrics. Let $B(TP_j, FP_j, TN_j, FN_j)$ represent some specific binary classification metric ($B \in \{Accuracy, Precision, Recall, F1\}^4$). Finally, we used the four metrics for each label to calculate the micro-average metrics for the model using, based on:

$$B_{\text{micro}}(h) = B\left(\sum_{j=1}^{q} TP_j, \sum_{j=1}^{q} FP_j, \sum_{j=1}^{q} TN_j, \sum_{j=1}^{q} FN_j\right) \tag{2}$$

The results, as shown in Table II, include the outcomes of the extra six models we selected, which performed the same task for comparison.

In addition, we plotted the confusion matrices for some of the label classification results. Fig. 3 shows the classification outcomes for the seven symptoms in the patient questionnaire section of MDS-UPDRS Part I. It can be observed that the classification achieved a high accuracy.

V. Conclusions and Future Work

In this study, we present an innovative application, which uses LLMs to construct self-reported corpora for training data,

TABLE II: Comparing with existing methods on the MDS-UPDRS task.

Models	Accuracy	Precision	Recall	F1
BERT with Pretrain-MLP	0.2597	0.3615	0.3220	0.3406
GPT-3.5 (zero-shot)	0.7186	0.5604	0.8242	0.6671
GPT-3.5 (few-shot)	0.7492	0.6081	0.8432	0.7066
Fintuned-ChatGLM3-6B	0.8405	0.6014	0.9004	0.7211
GLM-4-Plus (zero-shot)	0.8840	0.7102	0.9275	0.8045
GLM-4-Plus (few-shot)	0.8936	0.7341	**0.9335**	0.8218
Ours	**0.9536**	**0.9085**	0.8543	**0.8805**

and employs a pre-trained medical BERT model for fine-tuning to better perform the daily classification tasks of the MDS-UPDRS. It validates the possibility of using artificial intelligence to report the results of Parkinson's disease scale in non-outpatient settings. Our system also demonstrates the potential for further continuous assessment and tracking of Parkinson's patients' conditions. In addition, compared to the current LLMs with massive parameters, our system is more lightweight. As an auxiliary diagnostic tool, it is also easier to deploy. For future work, we intend to investigate the utilization of the model's generated results for Parkinson's disease and examine more user-friendly interaction approaches for both patients and doctors.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 62371118.

REFERENCES

[1] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[2] E. Bolton, A. Venigalla, M. Yasunaga, D. Hall, B. Xiong, T. Lee, R. Daneshjou, J. Frankle, P. Liang, M. Carbin *et al.*, "Biomedlm: A 2.7 b parameter language model trained on biomedical text," *arXiv preprint arXiv:2403.18421*, 2024.

[3] H. Xiong, S. Wang, Y. Zhu, Z. Zhao, Y. Liu, L. Huang, Q. Wang, and D. Shen, "Doctorglm: Fine-tuning your chinese doctor is not a herculean task," *arXiv preprint arXiv:2304.01097*, 2023.

[4] L. Yang, S. Xu, A. Sellergren, T. Kohlberger, Y. Zhou, I. Ktena, A. Kiraly, F. Ahmed, F. Hormozdiari, T. Jaroensri *et al.*, "Advancing multimodal medical capabilities of gemini," *arXiv preprint arXiv:2405.03162*, 2024.

[5] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[6] J. D. Steinmetz, K. M. Seeher, N. Schiess, E. Nichols, B. Cao, C. Servili, V. Cavallera, E. Cousin, H. Hagins, M. E. Moberg *et al.*, "Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021," *The Lancet Neurology*, vol. 23, no. 4, pp. 344–381, 2024.

[7] J. Zhu, Y. Cui, J. Zhang, R. Yan, D. Su, D. Zhao, A. Wang, and T. Feng, "Temporal trends in the prevalence of parkinson's disease from 1980 to 2023: a systematic review and meta-analysis," *The Lancet Healthy Longevity*, vol. 5, no. 7, pp. e464–e479, 2024.

[8] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[9] R. Rajan, L. Brennan, B. R. Bloem, N. Dahodwala, J. Gardner, J. G. Goldman, D. A. Grimes, R. Iansek, N. Kovács, J. McGinley *et al.*, "Integrated care in parkinson's disease: a systematic review and meta-analysis," *Movement Disorders*, vol. 35, no. 9, pp. 1509–1531, 2020.

[10] G. Wang, X. Liu, Z. Ying, G. Yang, Z. Chen, Z. Liu, M. Zhang, H. Yan, Y. Lu, Y. Gao *et al.*, "Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial," *Nature Medicine*, vol. 29, no. 10, pp. 2633–2642, 2023.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2024.

[13] T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, H. Wang, J. Sun, J. Zhang, J. Cheng, J. Gui, J. Tang, J. Zhang, J. Li, L. Zhao, L. Wu, L. Zhong, M. Liu, M. Huang, P. Zhang, Q. Zheng, R. Lu, S. Duan, S. Zhang, S. Cao, S. Yang, W. L. Tam, W. Zhao, X. Liu, X. Xia, X. Zhang, X. Gu, X. Lv, X. Liu, X. Liu, X. Yang, X. Song, X. Zhang, Y. An, Y. Xu, Y. Niu, Y. Yang, Y. Li, Y. Bai, Y. Dong, Z. Qi, Z. Wang, Z. Yang, Z. Du, Z. Hou, and Z. Wang, "Chatglm: A family of large language models from glm-130b to glm-4 all tools," 2024.

[14] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ digital medicine*, vol. 4, no. 1, p. 86, 2021.

[15] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[16] 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*, vol. 15, no. 6, 2023.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[18] W. Rahman, A. Abdelkader, S. Lee, P. Yang, M. S. Islam, T. Adnan, M. Hasan, E. Wagner, S. Park, E. R. Dorsey, C. Schwartz, K. Jaffe, and E. Hoque, "A user-centered framework to empower people with parkinson's disease," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 7, no. 4, Jan. 2024. [Online]. Available: https://doi.org/10.1145/3631430

[19] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury *et al.*, "The parkinson progression marker initiative (ppmi)," *Progress in neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2024.

[21] P. J. Lavrakas, *Encyclopedia of survey research methods*. Sage publications, 2008.

[22] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.