Group 9 – BDSA Project

Theme – #5 Generative AI agents: Comparative Analysis of Diagnostic Data with Modern Pre-trained Large Language Models

Aishwarya Solanki, Nikhil Sachdeva, Swapna Gadre, Bharat Raghunathan, Sidhant Subramanian

Motivation and Objectives

Over the last century, the average lifespan of a human has increased from 40 years to 72 years due to the advancements in healthcare which have helped us prevent diseases, diagnose early onset of many illnesses, and seek effective treatment. For an instance, the COVID-19 vaccine was developed and distributed in a much shorter timespan than the vaccines for pandemics of the past, thanks to the progress in healthcare technology. In recent years, mental health is also gaining much-needed attention which has improved the daily well-being of individuals.

Despite these advancements, many surgical procedures still rely on human control in real-time, even as robots have begun to outperform humans in certain areas, such as neurosurgery and ENT surgeries for tumor removal. Moreover, while healthcare has made significant strides, access remains a challenge; it is not universally available, and even when accessible, it is often not affordable.

With the increased availability and access of open-source models from Tech Giants like Facebook, Google, and Open-AI, along with advanced prompting techniques, Researchers are exploring whether these Large Language Models (LLMs) can be used to automate pre-treatment processes in healthcare, like identifying diseases in humans based on symptoms they feel and provide at least some basic remedies to help relieve pain in times when doctors cannot be accessed. Large language models (LLMs) are advanced artificial intelligence systems designed to understand and generate human-like text based on vast amounts of data. They utilize deep learning techniques, particularly neural networks, to analyze language patterns and produce coherent responses across various contexts. LLM's are trained on the broadest knowledge base and are extremely fast in reading and analyzing data. They are also constantly updated with new information and are multilingual.

Related work

Use of LLMs in healthcare has increased exponentially, especially since OpenAI publicized their own ChatGPT-3, which was treated by the public users as a groundbreaking product for everyone to use.

There's plenty of studies on utilizing the potential of the generative capabilities of LLMs, with some of them focusing on the healthcare domain. One study in particular which focuses on how such models can be leveraged in clinical settings was done by Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. *et al.*, where they discussed the strengths and limitations of LLM usage, and how chatbots are being created to leverage the latest technologies available to the public in aiding medical providers. [6]

There are a lot more studies which experiment with various models and architectures trying to understand how to maximize the potential for medical assistance using LLMs. In one study by Yang, X., Chen, A., PourNejatian, N. *et al*, the authors developed a large clinical language model GatorTron using 90B tokens, including 82B words of de-identified clinical texts. This study was primarily aimed at experimenting with the scaling of model parameters and the scope of training data and evaluating how these factors affect the accuracy of the model. The evaluation was carried out on 5 different language

processing tasks - inference, semantic similarity, medical answering, relation extraction, and concept extraction. [4] A follow up study by some of the same authors created a generative clinical LLM called GatorTronGPT, which was developed based on a GPT-3 architecture, using 277B words of text, this time adding 195B words of wide-ranging common English text to 82B words from the clinical texts. [3]

While in general LLMs have been very successful in democratizing knowledge, there's a lot of concern for, and a few studies investigating, the potential for spreading misinformation far and wide. Such problems are a huge risk to the public, and hence, it is very important to inject some accountability and clearly establish the limits of LLMs when it comes to medical advice. One such study was conducted by Clusmann, J., Kolbinger, F.R., Muti, H.S. *et al.*, providing a comprehensive overview of the potential and limitations of LLMs in healthcare. [5] There's even some studies that focus solely on literature review for various LLM applications in medicine, like one by Meng, X., Yan, X., Zhang, K. *et al.* where the authors primarily focus on the advances being made in medical writing comprehension, diagnostics, and the like, and suggest that future research should focus on deeper algorithmic understandings and ensuring responsible and safe AI usage. [7]

Proposed work

The project aims to determine the optimal set of prompts for various LLMs, including but not limited to BART, T5, GPT-4o, Llama, and Gemini, specifically for disease diagnosis tasks using symptom descriptions. This comparative study will further validate how the performance of the models will vary with different prompt engineering techniques to identify which types of prompts will work best for each of these models for accurately diagnosing medical conditions based on patient symptoms.

The innovative approach this project brings to the field of medical diagnosis using LLMs encompasses several elements. Firstly, by developing novel prompts specifically tailored for disease diagnosis tasks, the study aims to enhance the accuracy and relevance of AI-generated diagnoses. This will be further complemented by a unique comparative analysis of the diagnostic capabilities of various LLMs using these diverse prompting strategies, carving out new insights into strengths and weaknesses of different models. The focus of creating prompts which can stand in real-world clinical application strengthens the commitment this project has for developing AI tools that could assist real-world clinical decision-making and improve patient care and diagnostic accuracy in healthcare settings. Lastly, from the symptom dataset that we gather, we will apply various prompting techniques like Zero-shot prompting, One-shot prompting, Few-shot prompting, Chain-of-thought prompting, Tree-of-thought prompting to the Large Language Models (LLMs) and analyze the outputs. These techniques are elaborated in the "Evaluation and Testing Methods" section later in the proposal. An additional thing to note here is that, using this project, we will also measure the response a LLM model outputs to symptoms of depression, anxiety, and other known mental disorders. In these cases, along with a diagnosis, we aim to check the LLM's conversational abilities and how responsibly they are trained to answer such prompts.

Plan of action

Resources required:

1. **Software**: Access to the Large Language Models (LLMs) as BART, T5, GPT-4o, Llama, and Gemini via their respective APIs or frameworks. This list of models is not exhaustive and as we proceed with the project, we might amend this list to include models that our research suggests performing better on healthcare data or some healthcare-specific models that show promising

results which can be used for comparative analysis. Additionally, we might use Python libraries like OpenAI for evaluating performance on various parameter settings with different prompts, Hugging Face Transformers for model implementation and evaluation, and basic tools like IDEs or Jupyter Notebook for data analysis and visualization.

2. **Hardware**: A modern laptop which can run LLM APIs efficiently should suffice in this case. Additionally, we can leverage cloud computing resources (e.g., AWS or Google Cloud) to ensure sufficient computational power for running multiple models and processing large datasets.
3. **Dataset**: As discussed with professor, since LLMs are pre-trained on data mostly accessible and available on the internet, we might not need to re-train any model for comparative analysis. Rather, we can pick and select dataset for symptoms human being suffer when ill and prompt the model to predict what the disease could be. For comparison purposes, we will have a correct answer that will be checked against all the model outputs to gain insights into how efficiently a model performs.

Schedule and Milestones:

Week 6 of classes:

1. Research on the topic and finalize the Project proposal.

Week 7:

1. Research and Finalize the LLM models to be used and divide the work among members.
2. Gather dataset and setup testing environments.
3. Present the project idea to Professor and gather feedback.

Week 8:

1. Implement the feedback received and update the project.
2. Develop baseline models using zero-shot and one-shot prompting techniques.
3. Begin initial model evaluations and document results.

Week 9:

1. Implement few-shot prompting and chain-of-thought prompting techniques.
2. Document results and compare performance metrics across models with various hyper-parameter settings.

Week 10:

1. Explore tree-of-thought prompting methods and refine prompts based on previous evaluations.
2. Conduct thorough testing and validation of all models for a single prompt and an overall analysis of all the models at once.

Week 11-12:

1. Prepare a comprehensive report detailing methodologies, findings, and implications for healthcare applications.

Week 13:

1. Present the project idea and findings during the Workshop presentation.

Week 14,15:

1. Incorporate feedback, update project, and the final report to prepare for a final Demo.

Week 16:

1. Demo with Professor.

Evaluation and Testing Method

For instances in the dataset where we possess the true diagnosis, we can evaluate various evaluation metrics such as:

- **Precision:** Out of the diagnoses made by LLM, how many are correct, or stated differently, among the predicted diagnoses, how many are true positives.
- **Recall (Sensitivity)**: Refers to the number of actual diagnoses that are correctly predicted by LLM; it essentially refers to the capability of the model in catching all the relevant diagnoses.
- **Accuracy**: LLM diagnosis tallies with the actual diagnosis on the dataset.

In the case where we don't have the ground truth or the ideal/correct diagnosis, consistency metrics between the various models can be checked, such as:

- **Inter-model Agreement:** This would be a comparison between the consistency of the diagnoses across different LLMs. As an example, you could measure the percentage of overlap in predictions. This would also be a useful tool in catching instances of hallucinations in one of the models.
- **Intra-model stability**: Check whether the same LLM gives consistent diagnoses when slight variations in symptom descriptions, such as symptom rewording and reordering, are given.

Furthermore, depending on the complexity of the diagnoses and prompts in the dataset, some additional metrics to consider are:

- **Symptom Interpretation:** Assess the performance of each LLM in challenging cases, like multiple symptoms or symptom complexes requiring subtle understanding, such as when many diseases share several symptoms.
- **Handling Rare Diseases:** Assess the capability of LLMs in identifying rare diseases where large training datasets may not be available.

The types of prompting methods that will be used for the evaluation of the various models are:

- **Zero-Shot Prompting:** In the context of this dataset, this will involve listing the symptoms and asking the model for the diagnosis.
- **One-Shot Prompting:** In this dataset, it translates to one example of the correct diagnosis given a symptoms list, then listing the symptoms and asking the model for the diagnosis.
- **Few-Shot Prompting:** Like one-shot but giving multiple examples of different diagnoses given different symptoms lists, then giving a list of symptoms for which one wants the model to produce the diagnosis.
- **Chain-of-Thoughts Prompting (CoT):** Presented in the paper (Wei et al., 2022), the findings are that LLMs produce better, more accurate responses to tasks when the reasoning taken by a human to arrive at the answer is explicitly given in the prompt. For this dataset, this would amount to explicitly noting the reasoning taken by an actual human doctor in ruling out certain diagnoses, favouring certain diagnoses and finally predicting the closest one given the list of symptoms.
- **Tree-of-Thoughts Prompting (ToT)**: Established by (Yao et al., 2023), this approach takes it a notch more than CoT, in that it decomposes the thoughts into steps and generate candidates for

each intermediate thought step. For example, in the context of the symptom dataset, one step would be narrowing down the region of the body where the symptom is present, another step or branch of thought would be prioritizing common conditions vs further inquiry/investigation into the patient's demographic or medical history to ascertain any rarer conditions etc. In general, it decomposes the thought process of finding a diagnosis into steps, with multiple candidates or branches for each step.

Each of the models will be evaluated against the same set of prompts (at least one prompt for each of the different types mentioned above) and the classification metrics will be evaluated wherever the ground truth (ideal diagnosis or human diagnosis) are available and the results obtained will be used to compare and contrast which model is better suited for diagnosis prediction under a particular set of conditions.

In summary, this project seeks to harness the potential of Large Language Models (LLMs) to compare disease diagnosis through various prompting techniques. By systematically comparing the performance of various models—such as BART, T5, GPT-4o, Llama, Gemini, and more—using tailored prompts for symptom analysis. Our findings will not only contribute to the evolving landscape of AI in healthcare but also strive to ensure that these technologies are accessible and responsible, ultimately improving patient care and diagnostic accuracy in clinical settings.

Bibliography

1) Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.11903

2) Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2305.10601

3) Peng, C., Yang, X., Chen, A. *et al.* A study of generative large language model for medical research and healthcare. *npj Digit. Med.* **6**, 210 (2023). https://doi.org/10.1038/s41746-023-00958-w

4) Yang, X., Chen, A., PourNejatian, N. *et al.* A large language model for electronic health records. *npj Digit. Med.* **5**, 194 (2022). https://doi.org/10.1038/s41746-022-00742-2

5) Clusmann, J., Kolbinger, F.R., Muti, H.S. *et al.* The future landscape of large language models in medicine. *Commun Med* **3**, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1

6) Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K. *et al.* Large language models in medicine. *Nat Med* **29**, 1930–1940 (2023). https://doi.org/10.1038/s41591-023-02448-8

7) Meng, X., Yan, X., Zhang, K. et al. The application of large language models in medicine: A scoping review. iScience, 27(5), 109713 (2024). https://doi.org/10.1016/j.isci.2024.109713