PharMe: A Pharmaceutical Informed LLM

Keeping Healthcare Providers Up to Date with the Latest Treatments

Dan Harvey
Computer Science
Columbia University
dyh2111@columbia.edu

Berk Yilmaz
Electrical Engineering
Columbia University
by2385@columbia.edu

Ishraq Khan
Data Science
Columbia University
imk2133@columbia.edu

***Abstract*—This study explored the application of Medical Llama-3-8B, a large language model pre-trained on the MIMIC-III dataset, as a resource tool for assisting healthcare providers in drug and treatment selection based on patient diagnoses. The model was fine-tuned on data sourced (324 Updates for 2024) from Drugs@FDA, specifically the FDA's Purple Book database, which provides comprehensive information on approved drugs, novel treatments, and biosimilars.**

To ensure the model remained current with the latest updates, an automated workflow was developed using Apache Airflow. This workflow facilitated periodic data pulls from the FDA database, processed and formatted the retrieved information, and incorporated it into the fine-tuning loop.

The resulting fine-tuned model, termed PharMe, demonstrated the ability to provide valuable recommendations when prompted with medical conditions, achieving a perplexity score on average of 17.4, compared to GPT-Neo's score of 21.7. PharMe not only suggested commonly prescribed treatments but also identified the latest advancements, including novel therapies, and biosimilars.

This project demonstrated the potential of large language models when fine-tuned for specific applications in the field of medicine and serves as a promising example of how such models can support healthcare decision-making and leveraging AI in the medical industry.

I. Introduction

The rapid pace of pharmaceutical innovation and regulatory updates presents an unprecedented challenge in modern healthcare delivery. Providers are tasked with providing quality care in a timely manner, while simultaneously processing mountains of reporting and administrative duties. This makes it increasingly difficult to stay on top of novel drug approvals and existing drug modifications released by the Food and Drug Administration (FDA) on a monthly basis. In 2024 alone, the Food and Drug Administration (FDA) approved 44 novel drugs and over 10,000 updates to existing drugs [1]. With potential implications on patient care management and regulatory compliance, it is necessary to find a way to integrate novel drug information into clinical practice in a streamlined fashion.

Healthcare providers are solely responsible for determining the most appropriate treatments for their patients, considering factors such as the current diagnosis, comorbidities, previous medical conditions, allergies, and preferences for brand-name drugs or biosimilars due to insurance coverage. Additionally, they must account for potential contraindications and incompatibilities, especially when a patient is undergoing multiple treatments. While modern electronic health record (EHR) systems can prepopulate drug options and flag critical issues, the responsibility for staying informed about the latest treatments ultimately lies with the providers.

Healthcare providers are frequently visited by medical sales representatives whose primary role is to present information about their companies' products and services. In addition to providing educational materials, med reps often offer incentives such as department-wide meals or sponsorships for trips and conferences. While these interactions can help providers stay informed about new treatments and innovations, they may also introduce biases, as the information presented typically focuses on the products being promoted by the rep's company.

One common method for staying updated on medical advancements is subscribing to relevant journals within one's field. While this approach is more accessible and cost-effective, it places the burden of filtering through advancements on the provider. Furthermore, journal publications can sometimes be influenced by sponsorships from manufacturers, which may lead to selective reporting, leaving out updates from less-prominent companies or competitors. Additionally, physical journals are considered long lead outlets, requiring several months for research, writing, editing, and printing. As a result, the information may already be outdated by the time it reaches providers, typically lagging 2-3 months behind current developments.

Another avenue for staying current is attending annual conferences, which provide opportunities to learn about the latest research, treatments, and technologies. However, these multi-day events are often costly and rely on financial support from medical sales representatives or employers. Conferences also tend to favor larger pharmaceutical and medical manufacturing companies with the resources to sponsor prominent displays and presentations, potentially overshadowing smaller, less funded organizations with novel contributions.

In private practice, the challenge of staying up-to-date becomes even more demanding. Providers in these settings must take on the full responsibility of continuing education and staying informed about advancements in their field. This not only fulfills their professional obligations as practitioners but also reflects their commitment to delivering the highest standard of care to their patients. Balancing these demands with clinical responsibilities underscores the complexity of maintaining medical expertise in a rapidly evolving healthcare landscape.

This project aims to address these challenges by developing a clinically trained, condition-aware large language model that incorporates continual updates from the FDA. The result is a real-time, dynamically updated platform designed to support clinical decision-making. This tool empowers healthcare providers by highlighting treatment options they might have overlooked, forgotten, or been unaware of, including cost-effective biosimilars and generics. By doing so, it enhances prescribing accuracy and cost-efficiency, ultimately improving patient care.

II. Related Works

Several studies have explored the application of large language models (LLMs) in various facets of healthcare and the pharmaceutical industry. One study investigated the use of LLMs in drug recommendation systems [2], evaluating models such as RETAIN, G-Bert, GAMENet, SafeDrug, MICRON, COGNet, and REFINE. These models leveraged patient data from EHRs to generate drug recommendations. However, challenges emerged in making recommendations for new patients with limited or no prior medical history. The study's best-performing model, LEADER, demonstrated superior accuracy and efficiency compared to other models, yet emphasized the need for further research to incorporate drug interactions, enhancing safety and precision.

Another comprehensive study analyzed 7,402 research papers to explore the broader applications of LLMs across the pharmaceutical sector, including research and development (R&D), manufacturing, quality control, regulatory affairs, and clinical settings [3]. Similar to PharMe, this study examined the use of chatbots as consultants during surgical procedures, where natural language processing (NLP) was employed to assist surgeons in critical decision-making. These chatbots provided real-time recommendations based on clinical guidelines and medical literature. Additionally, the study highlighted the development of GA-DRUG, a genetic-awareness model designed to predict drug interactions by analyzing individual patient genetic profiles, thereby enabling personalized treatment plans. Other models examined in this research acted as diagnostic decision-support tools, analyzing patient data, generating insights, and flagging anomalies. While these systems align with PharMe's goals, their scope is broader and less focused on referencing specific drugs for particular medical conditions.

A separate study delved into the applications of LLMs in the pharmaceutical supply chain [4]. This research underscored the potential of LLMs for inventory optimization, process automation, and systems integration. However, it also cautioned against risks such as AI hallucinations, where incorrect or misleading data could compromise operational integrity. The study emphasized that LLMs should complement human expertise rather than replace it, ensuring that decision-making remains robust and reliable.

Interestingly, none of the aforementioned studies have fully addressed the challenges or opportunities posed by biosimilars and generics. These areas are critical in modern pharmaceutical practice due to their potential to reduce costs and increase accessibility depending on the patient's health insurance or lack thereof. LLMs, with their ability to analyze large datasets and extract insights, could play a pivotal role in identifying the most cost-effective options among biosimilars and generics while ensuring they align with a patient's specific needs and genetic profile.

What makes this project unique is that PharMe focuses specifically on referencing applicable drugs for medical conditions, going beyond the functionality of a simple dictionary or database. Unlike traditional systems, PharMe considers the contextual nuances of the medical condition and integrates this understanding into its recommendations. Furthermore, PharMe is not designed as a diagnostic tool. Instead, it relies on healthcare providers to make the final prescribing decisions, alleviating many of the ethical and practical concerns associated with other medical AI systems. This targeted, context-aware approach sets PharMe apart as a safer and more focused application of LLM technology in healthcare.

III. Data

Data for this project includes two sources: one used indirectly (the MIMIC-III clinical database) and the other used directly (the FDA's drug database).

A. *MIMIC-III: A freely accessible critical care database*

MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely-available dataset released by Meta that is composed of de-identified electronic health record (EHR) data from over 40,000 ICU patients with admissions at Beth Israel Deaconess Medical Center in Boston, MA between 2001 and 2012 [5]. The dataset includes more than 40 million rows representing clinical data such as patient demographics, medications, diagnoses, vitals, and lab results. Although our team did not directly use this dataset during the training process, because of its importance to the fine-tuning process for the open-sourced Medical LLaMA 3B-8B model, it felt necessary to include an overview.

B. *Drugs@FDA: FDA Approved Drugs*

The dataset used directly in order to finetune our large language models comes from the FDA in two forms.

The first source is a consolidated list of FDA-Approved Biosimilar Products available for export on https://www.fda.gov/drugs/biosimilars/biosimilar-product-information. As one may expect, this is a relatively small dataset given the extensive process new pharmaceuticals undergo in order to be approved for public use. As of December 2024, there are only 63 FDA-approved biosimilars listed on the FDA website linked above. Steqeyma, the most recent addition to this list, is a biosimilar for Stelara, a medication used to treat certain autoimmune conditions, particularly psoriasis and inflammatory bowel diseases.

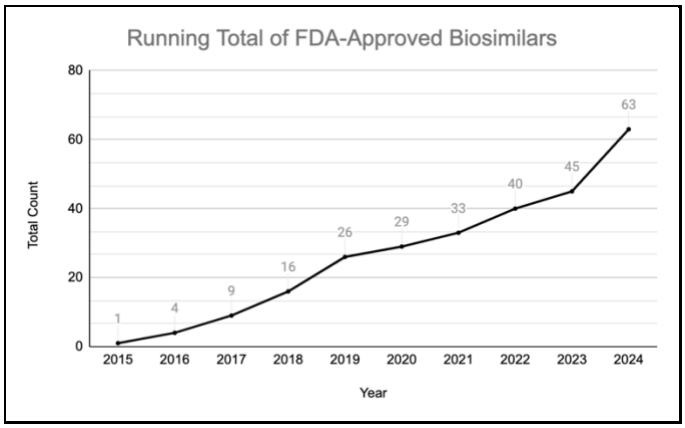

Fig. 1. Running total of approved biosimilars from 2015 to 2024

The second source is the FDA's Purple Book, an official publication that provides detailed information on 1) all FDA-licensed biological products regulated by the Center for Drug Evaluation and Research (CDER) and 2) all FDA-licensed allergenic, cellular and gene therapy, hematologic, and vaccine products by the Center for Biologics Evaluation and Research (CBER). Purple Book Data is available on-demand via the FDA's Drugs@FDA API and also in a monthly report uploaded directly to the FDA website as a CSV.

For our project, we decided to leverage the monthly report uploaded to the FDA website due to the consistently recurring nature of the uploads. Each report is divided into 2 sections:

1. The top section reports the changes made during the month in 3 categories.
    a) N: Newly approved products that received their initial FDA approval
    b) R: Products added in the current release which were previously approved products now included in the database
    c) U: FDA-approved products with changes to their information
2. The bottom section of the report provides a comprehensive list of all products in the Purple Book database for that month, including those listed in the top section that were either added or updated. This ensures that the report reflects the full scope of changes and new inclusions for that month.

We performed an analysis of the data available for 2024 and compiled our findings into the following graphs:

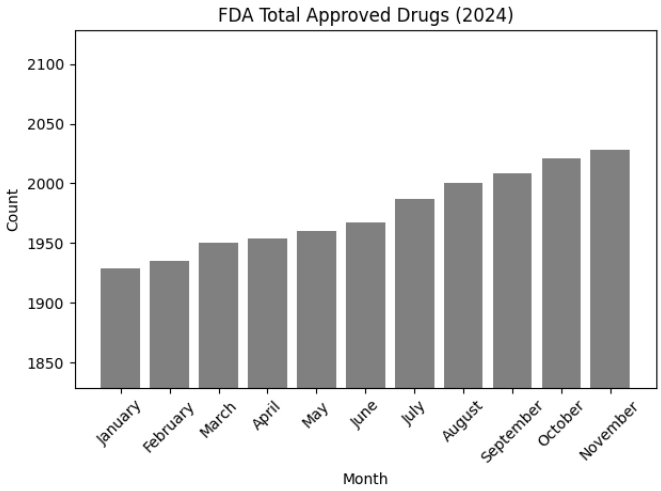

Fig. 2. Count of total FDA-licensed biological products in 2024

The number of total approved drugs grew from 1929 to 2028, an increase of 5.14%. This figure is only projected to increase at an even more rapid rate given the introduction of artificial intelligence into the drug discovery process.

Below, we see the variability in the number of NRU updates over the course of the year, ranging from a low of 17 in May to a high of 79 in August. The irregularity of the number of these updates adds to the challenge of keeping current for healthcare providers.

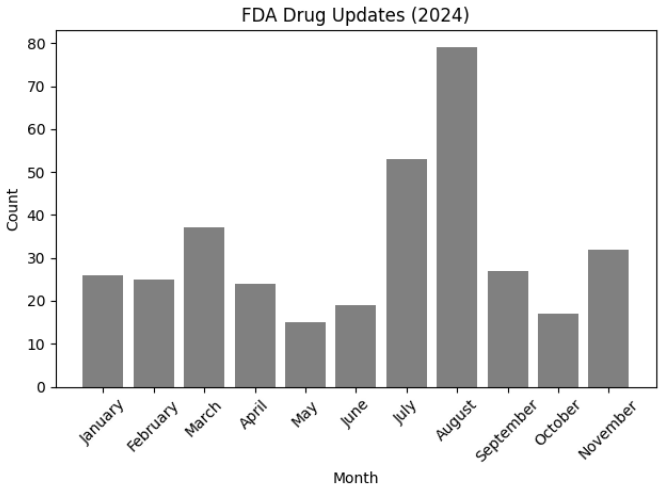

Fig. 3. Count of total changes made during each month in 2024

In terms of the actual data, the FDA provides structured CSVs which can be easily imported and processed as Pandas dataframes. The data includes information including product approval status, applicant, product details (e.g., name, strength, dosage form), marketing and licensure status, exclusivity expiration dates, and related reference product information. Due to the strict standards and regulations associated with biologics and pharmaceutical data, minimal preprocessing was necessary. Since not all of the data is relevant to our desired outcome (i.e. finetuning our clinical LLM to recommend biosimilars when providing clinical decision support based on a provided

diagnoses and symptoms), we carefully filtered our dataset to only the columns related to a product's identity, biosimilar status, and reference product information. A more detailed description of the dataset and its columns are provided below:

1. **NRU Date**: the year and month in which the NRU update for a specific product was released
2. **NRU**: the type of update ("N", "R", "U") for the biologic product
3. **Applicant**: the company that submitted the license application
4. **Proprietary Name**: the brand name of the biologic product
5. **Proper Name**: the generic name of the biologic product
6. **BLA Type**: the type of Biologics License Application (original, biosimilar, interchangeable)
7. **Approval Date**: the date the biologic product was approved by the FDA
8. **Ref. Product Proper Name**: the generic name of the reference biologic product for a biosimilar
9. **Ref. Product Proprietary Name**: the brand name of the reference biologic product for a biosimilar

IV. Methods

Our methodology for this project consists of several parts, including data collection and preprocessing using Apache Airflow, exploratory data analysis (EDA), feature selection, model finetuning, and output presentation. These different components are brought together to develop a fine-tuned clinical LLM capable of generating medication recommendations including the most recently approved biosimilars.

A. Data Collection

The data collection is a fundamental part of our methodology as it serves as the basis of the LLM finetuning process. We needed to decide between querying the Drugs@FDA API and scraping the Purple Book CSVs directly. Ultimately, the latter choice made more sense for our application given hardware limitations and the irregularity of NRU updates. Fine-tuning large models typically requires GPUs which we do not have unlimited access to as students. To find a balance between the most up-to-date data and cost-effectiveness, we decided to finetune on a monthly basis which aligns with how often the FDA releases its Purple Book CSVs. As a starting point, we collected all NRU updates from January 2024 to November 2024 (11 months) and consolidated this data into a single Pandas dataframe to be used for an initial round of fine-tuning.

The first step to aggregating data from all CSVs released in 2024 was to download the data. Ultimately we decided on building the URL manually as each CSV is uploaded directly to the fda.gov web server using a consistent format:

Fig. 4. Sample URL for January 2024 data

Because the CSVs are split into 2 different sections, we needed to parse out the relevant NRU data from the first section. We accomplished this by locating the index of both header rows in the CSV, converting the entire CSV into a Pandas dataframe, and then parsing the data between each index giving us the NRU dataframe. This process is repeated for every month of interest before we consolidate the results into a single dataframe representing the entire year. This final NRU dataframe had a shape of (354, 26).

B. Preprocessing and Feature Selection

With the NRU dataframe, we leveraged Pandas to perform all of our pre-processing for the initial fine-tuning. This mainly included column filtering for this initial implementation. Based on insights from our background research and domain knowledge, we conducted a manual feature selection.

C. LLM Fine Tuning

We first fine-tuned our model using the FDA API, creating a custom function to request HTTP to the FDA API for drug labels within a given specified date range. Then we extracted this data into Pandas dataframe to operate with. After that, with the help of our second custom function that creates biosimilar datasets, we extracted the biosimilar name and the reference product name to generate a dataframe to proceed with fine-tuning. To implement fine-tuning, we first created input-output pairs for the LLM model in the form of instruction which determines the reference product given a biosimilar name with the actual biosimilar name fetched from the dataset. We split our dataset into 90% training and 10% testing. As we have used in many coding sections for the initial model configuration, we used a 4-bit quantization configuration for efficient memory usage. After 50 epochs, we observed a training loss of 0.7213 and a validation loss of 0.7294.

D. Model Evaluation

One of the most crucial parts of this project was to compare different models and identify the best one based on selected evaluation metrics. The first of our metrics is perplexity, a measurement of how well a probabilistic model predicts a given input. The lower the perplexity, the higher the accuracy and prediction of the model outputs. We evaluated different models using a custom dataset that included 200 medications and their corresponding diseases. More details can be found under the Experiments section.

V. System Overview

The system architecture presents a comprehensive medical AI model pipeline that begins with an initial setup phase. This phase incorporates data from Drugs@FDA's Purple Book which undergoes data preprocessing and feature selection before being used to train a pre-trained Medical-LLaMA3-8B model.

The architecture then transitions into a recurring monthly operation phase, managed by Apache Airflow for automation. This phase maintains the same data sources and preprocessing pipeline but focuses on fine-tuning the existing model. The system culminates in a model output generation stage that incorporates additional input from symptoms & doctor diagnoses.

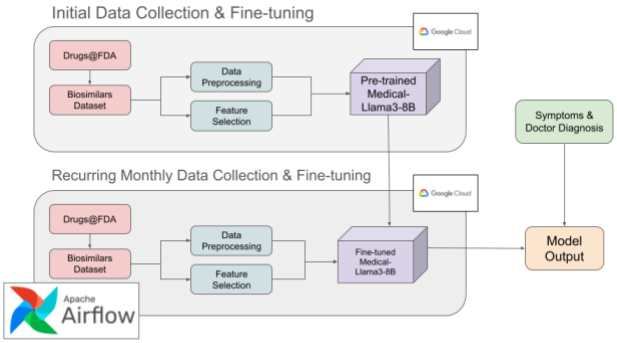

Fig. 5. System Architecture Diagram

A. *Airflow*

We use Apache Airflow, an open-source platform for workflow orchestration, to automate the process of parsing and processing monthly updates from the FDA Purple Book. Our workflow is showcased below. NOTE: not all portions have been fully implemented due to time constraints and hardware limitations related to LLM fine-tuning, specifically the finetune_llm task.

Fig. 6. Monthly recurring data collection & fine-tuning DAG

B. *Medical-LlaMa3-8b*

The main base architecture of LLM models is known as transformers. It was first described in the paper "Attention Is All You Need" by Vaswani et al [6]. Although some of the older methods in neural networks, such as RNN and LTSM, can utilize sequential data, they depend on the mechanism called "self-attention", to work with the different segments of the input rather than implementing it sequentially.

LLaMA3-8b is a large language model that uses the transformer-only architecture. By utilizing the tokens the model has seen before, the model predicts the next output by individually processing the text tokens. LLama3-8b has 8 billion parameters and implements the byte-pair encoding (BPE) and Sentence Piece-based tokenizer [6]. In the deep stacks of this model, we can observe the multi-head self-attention and feed-forward layers. Additionally, it has been previously trained on diverse text datasets which is essential for our project since we are delivering a medical chatbot system. The text has to be converted into a numeric form before entering the model and to achieve this, the first text is split into smaller sub-words called tokens. After that, the generated tokens are mapped to the corresponding embedded vector during the training process so similar embedded vectors end up with the tokens that represent words with similar context. Furthermore, positional embeddings help the model identify the tokens' positions [7]. LLaMA3-8b has been trained in pre-training and self-supervised learning to predict the next tokens using the context in the older ones [8]. In Figure 6, we can observe the general structure of the LLaMA3-8b model architecture. After the tokenization and embedding layer, 32 more layers follow this general structure before entering the final dense layer.

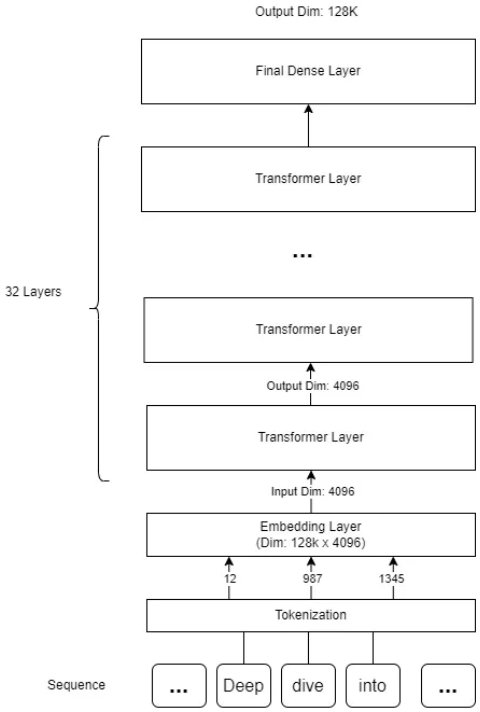

Fig. 7. Architecture of LLaMA3-8b [8]

Medical-LLaMA3-8b implements the same architecture as LLaMA3-8b, but is fine-tuned on several medical datasets, such as the AI Medical Chatbot dataset [9], LLaMA-3[8B]-MeditronV1.0, Bio-Medical-LLaMA-3-8B dataset, and a large number of medical and general instructions. Variants of this model have outperforme the state-of-art models in benchmarks such as MedQA and MedMCQA [10].

C. *Inference GUI*

The user interface for local implementation was developed using Tkinter, a Python GUI library. Tkinter provided a straightforward way to create a graphical interface for the model. However, its limitation is that it operates as a Python script, making it difficult to package into a standalone application like a .exe for Windows or .app for macOS for public deployment. Despite this, it offered a functional solution for local deployments.

For web-based implementations, the HTTP interface was built using HTML5 and CSS, providing a modern and responsive design. The API also allowed the creation of mobile web apps and desktop applications with web connectivity. These apps simply needed to make HTTP requests to the server hosting the model. Additionally, we recognized the potential of pseudo-

app structures—applications that appear native but are essentially built on web technologies like HTML and CSS. This approach could mimic native app behavior while leveraging the web's flexibility.

VI. Experiments

A. Data Collection

Initially, our approach to downloading data from the Drugs@FDA website involved using Selenium to scrape the URLs for the CSV files directly from the webpage. We planned to automate the navigation and retrieve the URLs from the page, then use Python's requests library to fetch the actual data. However, we encountered an issue where the page's JavaScript was not fully loading before the Selenium script attempted to extract the data. This likely happened because Selenium was not waiting long enough for the JavaScript-driven content to load dynamically, especially if the page used AJAX or other client-side rendering techniques. As a result, the request to fetch the CSV links didn't retrieve the full set of data, leading to incomplete or missing URLs.

After running into these issues, we explored a few different Python libraries designed to handle asynchronous loading of data, such as Playwright and Pyppeteer. However, after testing these options, we found that the most straightforward solution was to directly examine the structure of the URLs used to access the CSV files and see if we could manually generate them. By analyzing the webpage's network requests and identifying a consistent pattern in the URL structure, we were able to craft direct download links for the datasets. This approach bypassed the need for complex JavaScript rendering or asynchronous handling, allowing us to download the data directly and efficiently.

B. Model Evaluation

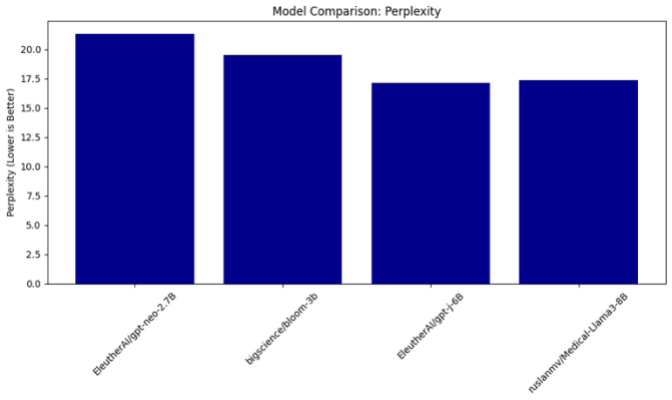

Fig. 8. Perplexity values for different LLM models

As can be seen in Figure 8, after the first evaluations, we observed that GPT-neo-2.7B had a perplexity of 21.3392 with a response time of 2.1917 seconds. The bloom-3b LLM model had a perplexity of 19.5437 and a response time of 1.5330. GPT-j-6B, on the other hand, had a perplexity of 17.1130 and a response time of 2.4248. Finally, the last compared LLM model, Medical-LLaMA3-8B, had a perplexity value of 17.3792 with a response time of 1.7368. When we compare the results, GPT-j-6B had the lowest perplexity level but a poor response time compared to the other models.

On the other hand, bloom-3b had the lowest response time; however, the perplexity values were higher than the models' average. Medical-LLaMA3-8B was the most stable LLM model, with a low perplexity value and a low response time. To conclude these evaluations, the first input was "Q: Suggest therapies for managing severe Systemic Lupus Erythematosus (SLE) flares."

Additionally, we wanted to observe the perplexity values with more complex diseases as can be seen in Figure 9.

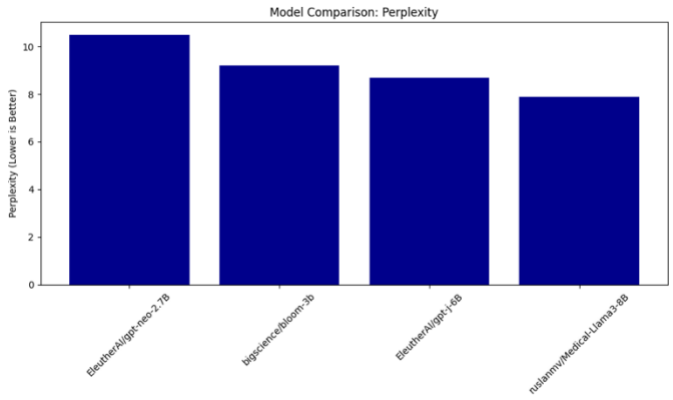

Fig. 9. Perplexity values for different LLM models

To achieve this, we changed the input sequence to "What medications are used for managing Mixed Connective Tissue Disease flares?" to capture differences in the perplexity values. As anticipated, we observed a drop in perplexity values when utilizing Medical-LLaMA3-8B. Following these evaluations, we compared the accuracy, prediction, and F1-score based on a dataset we created containing 200 medications and their corresponding diseases. Based on these evaluations, we selected Medical-LLaMA3-8B as our baseline model, achieving over 90% for each evaluation metric. After the evaluations, we initialized Medical LLaMA3-8b as our baseline model.

C. Interface

Interfacing with the model was designed to support inference locally or via an API, offering flexibility in its final deployment. The local implementation was the most straightforward to achieve. During the initial model buildout, we simply prompted the model using Jupyter Notebook, which provided a simple development environment. Similarly, hosting the model on a local HTTP server, such as Apache, enabled seamless interactions when both the server and client resided on the same virtual machine (VM).

To extend the model deployment options, we explored remote access by developing an API around it. Using a simple HTTP script with Flask or FastAPI libraries with authentication, we created a mechanism for submitting requests to the model and receiving the expected responses. This approach ensured secure and efficient remote access while maintaining the flexibility to integrate with various systems.

These design choices ensured maximum versatility in theory for delivering the model through various methods, whether as a local application, a remote API, or a hybrid web-based interface.

VII. Conclusion

This study investigated the feasibility and effectiveness of creating a state of the art, condition-aware large language model (LLM), nicknamed PharMe. Utilizing Medical LLaMA-3-8B, pre-tuned on the MIMIC-III dataset and further fine-tuned with data from the FDA's Drugs@FDA database and Purple Book, PharMe was able to create a solution for healthcare providers with actionable insights into traditional,cost-effective treatment options, including biosimilars and generics, that they may have otherwise overlooked or been unaware of, for any given diagnosis.

In conclusion, PharMe demonstrated the potential of large language models in healthcare by addressing potential gaps in pharmaceutical knowledge dissemination. By combining some of the latest models in AI technologies with healthcare domain-specific data, PharMe supports healthcare providers, not replacing them, in delivering informed, cost-effective, and high-quality care.

VIII. Future Work

Future work could focus on expanding the model's capabilities - specifically in condition awareness, enhancing its integration with electronic health record systems, and conducting further validation studies to ensure its applicability across diverse healthcare environments.

Another promising area for future research lies in tailoring treatment recommendations based on a patient's existing comorbidities, conditions, and current medications. By incorporating an advanced understanding of pharmacology and drug delivery mechanisms, the model could identify opportunities where a single medication might replace multiple drugs a patient is currently prescribed. This approach could streamline treatment regimens, reduce potential drug interactions, and improve overall patient outcomes.

REFERENCES

[1] U.S. Food and Drug Administration. (n.d.). Purple Book Search. U.S. Food and Drug Administration. Available: https://purplebooksearch.fda.gov/

[2] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Zijian Zhang, Feng Tian, Yefeng Zheng Large Language Model Distilling Medication Recommendation Model. Available: https://arxiv.org/abs/2402.02803

[3] Feng, Hui, et al. "Evaluation of large language model performance on the biomedical language understanding and reasoning benchmark: Comparative study. medRxiv." (2024): 2024-05. https://www.medrxiv.org/content/10.1101/2024.05.17.24307411v2.full.pdf

[4] David Aguero, Scott D. Nelson; The Potential Application of Large Language Models in Pharmaceutical Supply Chain Management. The Journal of Pediatric Pharmacology and Therapeutics 1 April 2024; 29 (2): 200–205. doi: https://doi.org/10.5863/1551-6776-29.2.200

[5] Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. https://doi.org/10.13026/C2XW26.

[6] Google AI, "SentencePiece: A simple and language-independent subword tokenizer and detokenizer," GitHub. [Online]. Available: https://github.com/google/sentencepiece. [Accessed: 20-Dec-2024].

[7] Meta AI, "LLaMA 3-8B Model Card," GitHub. [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[8] X. Zhao, "Deep Dive into LLaMA 3," Medium, Oct. 2024. [Online]. Available: https://medium.com/@zhao_xu/deep-dive-into-llama-3-351c7b4e7aa5.

[9] R. M. Villarroel, "AI Medical Chatbot Dataset," GitHub, 2024. [Online]. Available: https://github.com/ruslanmv/ai-medical-chatbot.

[10] H. Zhang, "LLaMA-3[8B]-MeditronV1.0: Advancing Medical AI Benchmarks," Hugging Face, 2024. [Online]. Available: https://huggingface.co/ContactDoctor/LLaMA-3-8B-MeditronV1.0.