ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/premkumar-ganesan/304

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 2 & Iss: 2

Research Article

LLM-Powered Observability Enhancing Monitoring and Diagnostics

Premkumar Ganesan*

Premkumar Ganesan, Technology Leader in Digital Transformation for Government and Public Sector, Baltimore, Maryland, USA

Citation: Ganesan P. LLM-Powered Observability Enhancing Monitoring and Diagnostics. *J Artif Intell Mach Learn & Data Sci 2024*, 2(2), 1329-1336. DOI: doi.org/10.51219/JAIMLD/premkumar-ganesan/304

Received: 02 May, 2024; **Accepted:** 18 May, 2024; **Published:** 20 May, 2024

***Corresponding author:** Premkumar Ganesan, Technology Leader in Digital Transformation for Government and Public Sector , Baltimore, Maryland, USA

ABSTRACT

Large Language Models (LLMs) are now essential for cutting-edge AI-powered applications. Full utilization of their potential and production of high-quality products can be achieved through effective observability, though. The ability to get detailed insights into how these complex models behave and perform is called LLM-observability. Through the systematic collection and analysis of logs, analytics and traces, businesses can gain a deeper insight into the internal workings of their LLMs. The fast development of LLMs like GPT-4, Gemini and GPT-3.5 has opened revolutionary possibilities in digital diagnostics and other fields of healthcare. Through symptom analysis and the identification of diagnoses that correspond well with prevalent illnesses, this study evaluates the diagnostic skills of each model and demonstrates how they might substantially enhance diagnostic accuracy and efficiency. Through a series of symptom-based diagnostic prompts derived from medical databases, GPT-4 demonstrates enhanced diagnostic accuracy, which is a result of its substantial training on medical data. Also, Gemini may be a reliable model for physicians to use when making potentially harmful diagnoses, given its excellent performance as a disease triage tool. A good diagnostic tool, GPT-3.5 isn't quite as state-of-the-art as GPT-3.6. This study highlights the importance of conducting more thorough research into LLMs for healthcare and clinical practices. It is crucial to ensure that any system using LLMs promotes patient privacy and complies with health information privacy laws, such as HIPAA compliance. Additionally, it is important to study the social consequences that impact diverse individuals in complex healthcare contexts. As the first in a series of studies, this one will look at how solving ethical issues with LLM's responsibility to learn from human biases can pave the way for new AI uses in complex healthcare contexts.

Keywords: LLMs, Healthcare, AI, Digital Health, Medical Diagnostics, Natural Language Processing (NLP), Machine Learning in Medicine

1. Introduction

Various aspects of healthcare, including interactions between doctors and patients, techniques of diagnosis and treatment, the importance of following treatment plans, changes in lifestyle or behavior and continuous preventative health efforts, all contribute to patients' overall healthcare experiences. The healthcare path of a patient is obviously not linear; rather, it is an extensive web of interconnected encounters and events[1]. Many factors, such as the patient's history, the nature of their symptoms, the clinician's level of training and experience, and the diagnostic instruments at their disposal, come together throughout the diagnostic process in medicine[2]. To aid in decision making, this data is subsequently analyzed for patterns. A treatment plan is developed, patient progress is monitored and disease development is tracked because of the process, which uses the acquired data to further refine and corroborate initial assumptions about the condition. To improve patient experiences while reducing the costs and workloads associated with primary care, recent innovations and continuing research have enabled substantial progress in digitizing a large component of the healthcare process. To streamline, automate and digitize healthcare operations, some approaches use ML, big data analysis and natural language processing (NLP)[3]. The automation of processes, the simplification of daily activities for all parties involved, the reduction of manual labor and the streamlining of workflows are all ways in which these technologies are expected to transform patient care and disease management[4]. The field of large language models (LLMs) is one example of an emerging technology that has the potential to completely alter the healthcare system. To be sure, LLMs show an impressive capacity to comprehend medical materials and recognize (diagnose) a wide variety of symptoms and illnesses. An excellent LLM is GPT by OpenAI, which powers ChatGPT and produces responses to text that are accurate and resemble human speech. One such prominent LLM is BERT (Bidirectional Encoder Representations from Transformers), developed by Google. Another is Llama (Large Language Model Meta AI), developed by Meta. Finally, Alpaca (fine-tuned from the Llama model), developed by Stanford. Although there are limitations to both LLM and NLP approaches, a combination of technologies can provide both effectiveness and cost-effectiveness.

Recent research has suggested a unique general three-step methodology to assess the utility of LLMs and ChatGPT, in healthcare diagnostics and therapy[6]. There has also been a review of ChatGPT's performance in terms of its communication capabilities in oncology and radiology. Depending on the situation, ChatGPT showed moderate to excellent performance. Additionally, it was shown that medical experts could improve ChatGPT's performance when working with other NLPs/ LLMs. This is because medical experts are better able to assess ChatGPT's responses than patients themselves.

What is LLM Observability?

Generative AI has rapidly grown increasingly vital to many areas of business, finance, security, research and language since the emergence of Large Language Models (LLMs) such as GPT, LaMDA, LLaMA and many more. To better manage frauds and boost conversion rates, Stripe partnered with Open AI and Microsoft introduced LLM-powered Copilot to enhance office productivity in common activities. Although there has been a marked increase in the use of LLMs, it has been more difficult to deploy LLM systems in production than regular ML apps. The challenge with LLMs comes from their large model sizes, complex architecture and non-deterministic results. The opaque nature of LLM applications' decision-making processes also makes debugging their associated faults a laborious and resource-intensive process. Maintaining LLMs' functionality and security through the generation of accurate and impartial responses necessitates constant monitoring. Teams may manage and understand the performance of LLM applications and language models with the help of LLM observability, which provides tools, approaches, and processes. This allows them to spot biases or drifts and fix problems before they affect the business or end-user experience.

What are the common issues with LLM applications?

The combination of artificial intelligence with LLM technologies is still in its early stages, thus there is room for improvement. Users and the LLM itself may encounter some challenges. With the right LLM monitoring technology, businesses can maintain tabs on potential problems with LLM applications, including:

Hallucinations: Applications powered by LLM can sometimes give you misleading information, also called "hallucinating," especially when they don't know the answer to a question. They often give answers that sound confident but are incorrect rather than admit they don't know anything. When using LLMs for tasks that require precise facts, it is essential to keep in mind that this tendency could encourage the dissemination of false information.

Efficiency and affordability: Many LLM-based applications depend on external models. Problems like third-party API performance drops, algorithm changes cause discrepancies, and excessive costs (particularly for big data) might result from being too reliant on them.

Prompt Hacking: Programs to generate predefined material. Inappropriate or damaging material may be generated by LLMs because of this manipulation. Being cognizant of this matter is crucial, especially when implementing LLMs in applications that interact with customers.

Privacy and security: LLMs present privacy and security concerns, including as the possibility of data leaks, biases in output from unbalanced training data, and the threat of illegal access. Also, LLMs could end up returning information that is private or sensitive. So, with LLMs, it's crucial to have strict security measures and ethical processes.

Variation in model prompt and response: LLMs receive a wide variety of user prompts with varying lengths, languages and levels of accuracy in their responses. Furthermore, users could get conflicting answers to the same question, which could cause them frustration and a lack of consistency in their experience. Because of this, keeping track of LLM applications and monitoring them constantly is essential.

2. Literature Review

The primary goal of this section is to lay forth all the necessary context for the methods that were employed. Here, we investigate natural language processing (NLP) from many angles and look at its many uses in healthcare. The healthcare industry has made great strides in integrating NLP and LLMs[7]. A growing number of applications are making use of these technologies. These include analyzing patient sentiment through reviews and feedback, extracting crucial data from electronic health records (EHR) and assisting decision making in clinical settings.

Natural Language Processing

NLP is an important branch of AI that studies how computers can understand and interact with human language. It makes several things easier, such as conversational interfaces, sentiment analysis and translation. Models like GPT and BERT are the result of NLP's long and winding road from rule-based approaches to complex ML techniques[8]. Key ideas in natural language processing include parsing, named entity recognition, tokenization and part-of-speech tagging. Understanding human language well, allowing for meaning extraction and reasoning simulation to complete tasks, is the goal. Natural language processing makes use of a wide variety of models and methods, from simple rule-based systems to complex ML algorithms. Among the most well-known models in natural language processing.

- Pattern-matching and substitution-based natural language processing: Earlier NLP systems relied on lexicons and rules that were hand-crafted. The ELIZA chatbot[9], which was built in 1964, is a famous example. One of the earliest programs that could attempt the Turing test was ELIZA.

- Text categorization and sentiment analysis often make use of ML models, which include more conventional models such as decision trees, naive Bayes and support vector machines (SVM).
- **Neural networks:** These models, which take their cues from the human brain, are great for jobs that need knowledge of the sequential structure of a language. Examples of neural networks are convolutional neural networks (CNNs) and recurrent neural networks (RNNs).
- **Embedding models:** These models capture semantic contents by producing dense vector representations of words or larger text units. Some notable examples include FastText, GloVe and Word2Vec[10].
- Machine translation and text summarization rely heavily on sequence-to-sequence models, which can convert input sequences into output sequences. These models often use an encoder-decoder architecture with attention mechanisms[11].
- The first type of natural language processing (NLP) model is the large language model (LLM), which can generate text, answer questions and translate between languages without specialized training data. What follows is a more in-depth discussion of LLMs.

Large Language Models (LLMs)

Autoregressive models used to generate text that is coherent within its context are part of the GPT series, which includes GPT-4 and GPT-3. With the help of language pattern understanding, GPT can decode the input and generate meaningful and consistent output. Examples of applications where GPT models shine include content creation, dialogue generation and other activities requiring the development of new text in response to supplied instructions. But BERT works by looking at the words that come before and after a sentence to get a better grasp on the context of the whole thing. Using a "attention" mechanism to give various words different levels of importance, the Transformer architecture-first proposed in[12]-forms the basis of both approaches. At the heart of these models is an encoder that takes word sequences and turns them into vector representations that are contextually enhanced. Word interdependencies over larger ranges can be included by these models thanks to their innovative self-attention mechanism, which greatly enhances their predictive accuracy. The use of a bidirectional training strategy allows BERT to predict words depending on both the previous and next context, which is very noteworthy. Contrast this with GPT's one-way approach. Before neural networks and Transformer models became the de facto standard in natural language processing, statistical models served as the backbone.

That which followed was crucial among them.

Markov Models: These probabilistic models take the present state as their sole determinant for the likelihood of each subsequent state, according to the principle named after mathematician Andrey Markov. Their use stands out in sequential tasks, such as language modelling.

Hidden Markov Models (HMMs): Hidden Markov Models (HMMs) offer both observable outputs and hidden states, making them an extension of Markov models. They are useful in natural language processing tasks, such as named entity recognition and part-of-speech tagging.

Conditional Random Fields (CRFs): To model the likelihood of outputs given inputs, natural language processing makes use of these statistical frameworks. More accurate results are produced by CRFs than by HMMs since they consider the complete word sequence.

n-gram Models: To forecast the subsequent item in a sequence, these models take into account the items that came before it (n − 1). Use of n-gram models, which assume that the likelihood of a word depends only on its preceding words, is widespread in domains such as machine translation and speech recognition.

Latent Dirichlet Allocation (LDA): (LDA) is a type of generative statistics that enables groups of unobserved variables to explain collections of observations. By viewing each document as a mixture of subjects and assigning each word to a document's topic, natural language processing (NLP) can discern why data portions are related.

Problems with NLP and Evaluation Pipeline

Below, we have outlined some significant issues and worries related to natural language processing (NLP), particularly as it pertains to its use in healthcare. The businesses who offer the most cutting-edge models for sale are actively working to solve some of these issues.

Hallucinations: The generation of results that have the appearance of plausibility but are actually completely false or manufactured.

Bias: The biases included in the training datasets are learnt and reproduced by LLMs[13]. Lack of explainability: Most of the time, generative AI systems won't explain their reasoning behind the results they produce or the responses they give[14]. Transparently representing the methods used to generate a response, categorisation or suggestion, Explainable AI (XAI) makes sure that consumers understand the properties of the used models. Another significant component that can promote transparency and lead to better acceptance of AI-empowered systems is considering user ability and providing personalization in XAI. The explainability of current GAI systems is lacking, especially when it comes to providing personalized explanations. Real-time validation: The data used to generate the answers is not up to date. Rather, they derive from the dataset that was utilized to train the model, which usually includes data over a time span that extends up until the tool's training date[15].

Limitations in mathematical operations: Python modules for calculations and more frequent model updates help alleviate this issue to some extent. Content-token size limitation: Token size limitations have been increased and usage prices have been raised to partially solve this limitation.

3. Methodology

3.1. Description of the LLMs Evaluated

Three accessible Large Language Models (LLMs) that are expected to be most valuable in clinical contexts-GPT-4, Gemini and GPT-3.5-are evaluated in this study for their diagnostic accuracy. The ability to generate new and clinically relevant information, as well as to make predictions and diagnoses in the clinical setting and to provide data-driven insights to support health maintenance and recovery, are all strengths of these models. These models require immediate assessment for their applicability and value in healthcare research and clinical tasks due to their widespread use and ease of access. There is potential for all three LLMs models to increase trust in medical diagnoses, as they each bring something special to the table when it comes to processing clinical statistics.

GPT-4: One of the most prominent entities in the field of language understanding and generation is GPT-4, developed by OpenAI. This model stands out due to its exceptional capacity to decipher intricate queries. It is worth mentioning that the architectural style is perfectly suited for evaluating the accuracy of diagnoses using descriptions of medical symptoms. With an impressive accuracy rate of 75% on the Medical Knowledge Self-Assessment Program, GPT-4 has proven to be highly effective in the medical field. Both the sophisticated interpretation of difficult medical queries and GPT-4's vital role in boosting diagnostic precision from symptom narratives are emphasized by this achievement.

Gemini: Gemini is a huge step forward for LLMs in general and for healthcare in particular thanks to its optimized architecture for domain-specific tasks. A great deal of care and attention went into designing Gemini so that it can better comprehend and generate complex answers in these niche areas. As a result, Gemini is a priceless asset for many endeavors, especially those requiring pinpoint accuracy, such as healthcare tests and investigations. A new standard for artificial intelligence in healthcare, its capacity to integrate and reason across multimodal inputs further highlights its potential to transform the processing and interpretation of medical information.

GPT-3.5: In terms of language understanding and generation, GPT-3.5, the predecessor of GPT-4, has considerable capabilities but is significantly less proficient than its successor. It serves as a foundational baseline. Its use provides a benchmark for assessing the development of LLMs and their potential use in healthcare diagnostics. Despite being an older version, GPT-3.5 has done great work in the medical field, with results like 53% accuracy on the Medical Knowledge Self-Assessment Program to brag about. This statistic highlights its capacity to handle and comprehend healthcare-related questions, which is a big deal for using AI to improve diagnostic precision.

3.2. Data collection methods

The primary data set used in this investigation was built using information retrieved from authoritative medical institutions such as the CDC, WHO, Mayo Clinic, Cleveland Clinic and Johns Hopkins Hospital. To assess the possible use of Large Language Models (LLMs) in offering diagnostic insights for common disorders, the disease selection criterion centered around problems often seen in daily life. For reasons of ethics and the present limitations of LLMs in accurately diagnosing chronic and complex diseases like cancer, we purposefully left out seasonal allergies, the common cold and food-related issues like diarrhea or allergies because of how common they are in the general population. A complete dataset was created by compiling a full list of symptoms for each chosen ailment and associating them with their respective names. To facilitate the development of diagnostic prompts, which aim to query illness predictions using symptom descriptions, this data was organized. And the Data Collection Process is shown in **figure 1**.

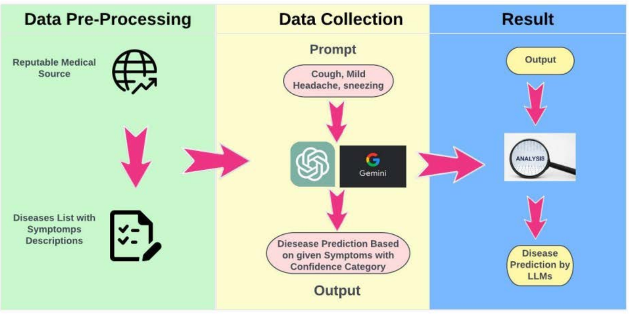

Data Pre-Processing
Data Collection
Result
Reputable Medical Source
Diseases List with Symptomps Descriptions
Prompt
Cough, Mild Headache, sneezing
Gemini
Diesease Prediction Based on given Symptoms with Confidence Category
Output
Output
ANALYSIS
Disease Prediction by LLMs

Figure 1: Data Collection Process.

The symptoms of each disease were painstakingly designed into these diagnostic prompts, which ask for disease predictions and a confidence score for each diagnosis. To ensure that the evaluation process was consistent, the prompts were administered consistently. The study's findings were supported by a strong methodological framework, which included manually verifying the replies to evaluate the accuracy and reliability of disease prediction. This approach highlights the study's goal of exploring the usefulness of LLMs as a tool to help people recognize prevalent health issues. Through an examination of common disorders, this research offers significant insights into the strengths and weaknesses of AI technology as they pertain to common health applications.

Prompt for models: To evaluate the diagnostic skills of different language models, the following dialogue provides a prompt.

3.3. Evaluation Metrics for Diagnosing Diseases through LLMs

A comprehensive, multi-stage manual procedure was used to assess the effectiveness of Language Learning Models (LLMs) in illness diagnosis based on medical symptom descriptions. Precision, recall and the F1 score-well-known metrics for providing a balanced view of predictive models' accuracy in detecting correct diagnoses and relevant omissions-were all used in this technique. We methodically classified each response according to its diagnostic accuracy as we examined the LLMs' outputs for every dataset item in our study. The categories were listed below:

- **True Positive (TP):** Examples of diseases that the LLM properly recognized, demonstrating how well the model can match symptoms with the right diagnosis.
- **False Positive (FP):** Cases where the LLM overestimated its diagnosis accuracy by wrongly diagnosing an illness based on symptom descriptions that did not correspond with the real sickness present.
- **False Negative (FN):** Cases where the LLM underestimated its diagnostic sensitivity because it either misidentified the patient's illness based on their symptoms or failed to identify any illness at all.

The following metrics were subsequently calculated:

- **Precision:** This statistic provides insight into the precision of the model's illness detection by measuring the exactness of its positive predictions, which are the fraction of TP observations among all positive diagnoses produced by the model.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** A measure of the model's thoroughness in disease identification, this metric evaluates its capacity to detect all relevant cases (i.e., the ratio of TP observations to all real positives inside the dataset).

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score:** This metric is useful when the contributions of recall and precision are about equal since it provides a balanced measurement of both. It provides a single metric for the model's total diagnostic performance and is computed as the harmonic mean of recall and precision.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Since it gives a fair assessment of both recall and precision, this statistic is helpful when their contributions are roughly equal. It is calculated as the harmonic mean of recall and precision and gives a single measure for the overall diagnostic performance of the model.

Evaluating LLMs

Verifying the precision of the model's predictions or outputs is the standard method for assessing a classic ML model. Standard measures like Accuracy, RMSE, AUC, Precision, Recall, and others are used to measure this. Assessing LLMs is significantly more intricate. Data scientists today employ a variety of methodologies.

1) Classification and Regression Metrics

It is simple to evaluate LLMs when they generate numerical predictions or classification labels. The process is identical to that of conventional ML models. We typically focus on assessing LLMs that generate text, albeit this can be useful in some situations.

2) Standalone text-based Metrics

In the absence of a ground truth source, these indicators are helpful for assessing LLM text output. You should consider academic recommendations, your own prior experience or the results of other models to establish what is considered appropriate. Confusion is one such case. The likelihood that the model would produce an input text sequence is measured, which may be seen as an evaluation of the model's learning performance on the training text. Additional instances encompass Reading Level and Non-letter Characters.

A more advanced method that can be used independently is to extract embeddings from the model's output and then examine them for any abnormal patterns. A 3D visualization displaying your embeddings' graph can be used to accomplish this manually. Your LLM application may be analyzed for bias and explainability by coloring or comparing by important areas like as gender, anticipated class or perplexity score. This might help you uncover any hidden flaws. In this method, embeddings can be visualized using a number of available software tools. Before mapping the embeddings into three dimensions, they group them together. While most use HDBSCAN and UMAP for this, a K-means-based technique has been employed by others. Along with visual evaluation, the embeddings can also be subjected to an anomaly detection algorithm that searches for outliers.

3) Evaluation Datasets

Textual output can be compared against a baseline of allowed responses using a dataset containing ground truth labels. As an example, the ROUGE metric is famous. When evaluating LLMs for language translation tasks, ROUGE compares their results to those of a reference dataset. By comparing results to a reference dataset, one may determine accuracy, relevance and many other criteria. One important part is embedded. You can compare the embeddings of your LLM output with the ground truth embeddings using standard distance metrics such as J-S Distance, Hellinger Distance, KS Distance and PSI. Finally, several LLM benchmark tests have gained widespread acceptance. For further information, check out Stanford's HELM page.

4) Evaluator LLMs

Using an LLM to assess another LLM may seem like cheating at first, but many see it as the way forward and research has showed positive results. I predict that in the not-too-distant future, Evaluator LLMs will reign supreme when it comes to LLM evaluation. The Toxicity measure is a classic example that everyone agrees on. The Toxicity of your model's output is determined by an Evaluator LLM. Hugging Face suggests using roberta-hate-speech-dynabench-r4. The results from the Evaluator LLM serve as the benchmark, hence all the measures mentioned in the section on Evaluation Datasets are applicable here. It is recommended that Evaluator LLMs be set up to give binary categorization labels for the metrics they evaluate, as per the findings of the Arise research. Binary labelling is more efficient and requires less effort than numerical ratings and ranking, scientists say.

5) Human Feedback

Manual, human-based input should not be overlooked amidst the focus on quantifiable measurements in this post, product documentation and marketing materials. When developing an LLM app, data scientists and engineers often think about this. An interface is typically included in LLM observability software to help with this. Incorporating human input into both the final review and continuing monitoring is recommended practice, just as it is with early development comments. You can learn a lot about your end result by collecting 50 to 100 input prompts and examining the output by hand.

Tracking LLMs

Prior to monitoring, there is tracking. There is sufficient complexity in the technicalities of tracking LLMs that my research justifies its own section. Accurately recording request volume, response time, token use, expenses and error rates is the easy fruit of tracking. While there are LLM-specific choices, there are also more general system monitoring tools that play a role here. The marketing departments of more conventional monitoring firms are eager to boast about their LLM Observability and Monitoring, which is based on tracking basic functional metrics. Recording input prompts and output responses allows for in-depth study in the future. Despite appearances, this is anything but easy. I, like most data scientists when discussing or writing about LLMs, have skipped over something complex so far. We will not be keeping tabs on, analyzing or evaluating an LLM. Here we have an application, which is a collection of components including one or more LLMs, agents and pre-set instruction prompts. Although not all LLM applications are very intricate, a growing number of them are. It can be challenging to determine the last prompt call in even moderately complex LLM applications. For debugging purposes, it is necessary to know the call's state and the order of its execution at each stage. Professionals in the field will benefit from using tools that simplify these intricacies.

Monitoring LLMs

Although the majority of LLMs and LLM applicants are evaluated in some way, not enough have instituted ongoing monitoring. We'll show you how to construct a monitoring program that safeguards your users and brand by dissecting the many parts of monitoring.

1) Functional Monitoring

The first step is to consistently keep an eye on the easy targets identified in the Tracking section. All of the following are part of it: request volume, response time, token consumption, expenses and mistake rates.

2) Monitoring Prompts

Paying attention to inputs or prompts provided by users should be your next priority. Readability and similar standalone metrics could provide useful information. Toxicology and similar issues should be addressed using evaluator LLMs. It is a good idea to incorporate embedding distances from the reference prompts as measurements. You still need to know if users are interacting with your app in novel ways, even if it can handle very different types of suggestions than you expected.

Now is the time to add a new assessment category: adversarial attempts, sometimes known as malicious prompt injections. The first assessment does not always take this into consideration. To identify malicious actors, it is possible to compare results to reference sets of known adversarial cues. Prompts can also be labelled as malicious or benign by evaluator LLMs.

3) Monitoring Responses

Several helpful tests can be put into place to compare the results that your LLM application is producing with your expectations. Think about how it relates. Do you get useful responses from your LLM, or is it all lost in thought? Is what you were expecting to be covered taking a different turn? Is sentiment the key? Are you getting the appropriate tone from your LLM, and is this altering as time goes on?

It is likely unnecessary to keep an eye on all these data every single day. For some, once a month or once every three months will be plenty. Worries about toxicity and hazardous output are, however, paramount whenever LLMs are used. Some metrics that you should monitor more frequently are these. The embedding visualization approaches we covered before might be useful for determining the source of an issue. The adversarial strategy of prompt leakage has not been implemented at this time. A prompt leakage happens when an unauthorized user can deceive your program into revealing the stored prompts. The time and effort you put into discovering which of the pre-set prompt instructions produced the best outcomes is evident. Personal information is at risk here. Respondent monitoring and comparison to your prompt instruction database can reveal prompt leaking. Metrics for embedding distance are effective. You should run your LLM application against evaluation or reference datasets on a regular basis and compare the results to see how it stacks up. Along with alerting you to drift, this can provide a feeling of accuracy over time. You can train your LLM to be more accurate on certain kinds of problematic prompts by exporting datasets of underperforming output, which is an option in various embedding management solutions.

4) Alerting and Thresholds

It is important to be cautious not to set your thresholds and alerts to trigger too many unnecessary notifications. It can be helpful to employ multivariate drift detection and alerting. How I want to accomplish this is something I will discuss in a later piece. By the way, in all the research I conducted for this piece, not a single source addressed top practices for thresholds or false alarm rates. It's discouraging.

For your must-have list, you might want to consider adding a few attractive alert-related features. You can connect your monitoring system to popular platforms like Slack and Pager Duty, which provide information streams. Input prompt alerts can trigger automatic response blocking in some monitoring systems. You can use the same capability to check if the response is free of toxicity, personally identifiable information leakage and other quality metrics before providing it to the user.

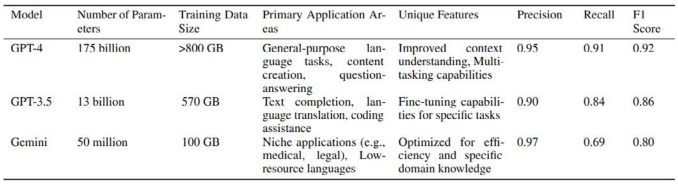
This is the best place for my last observation, so I'll leave it here. Depending on your monitoring method, custom metrics may be crucial. You might have something truly original in your LLM application or maybe a brilliant data scientist on your team came up with a statistic that would greatly improve your strategy. Innovation is expected to occur in this field. You should seek for custom metrics because of their adaptability.

5) The Monitoring UI

The ability to display measurements as time-series graphs in the user interface is a telltale sign that a system can be monitored. It's not that unusual. When user interfaces permit a level of root cause analysis by delving down into alarm trends, they begin to stand out. Some help with visualization of the embedding space using projections and clusters; I'd be interested in seeing or doing research on how useful these visualizations are in real-world scenarios. User, project and team monitoring will be consolidated in more advanced solutions. Based on the premise that all users are on a need-to-know basis, they will have RBAC. The fact that any user of the program has access to everyone's data is unacceptable to many modern businesses. The user interface does not allow for an adequate study of warnings, which is one reason why alerts often produce an unsatisfactory false alarm rate, as I said before. Even though it's not common, certain software systems do try to optimize in this way.

4. Results

Table 1: Model Performance and Characteristics.



Model	Number of Parameters	Training Data Size	Primary Application Areas	Unique Features	Precision	Recall	F1 Score
GPT-4	175 billion	>800 GB	General-purpose language tasks, content creation, question-answering	Improved context understanding, Multi-tasking capabilities	0.95	0.91	0.92
GPT-3.5	13 billion	570 GB	Text completion, language translation, coding assistance	Fine-tuning capabilities for specific tasks	0.90	0.84	0.86
Gemini	50 million	100 GB	Niche applications (e.g., medical, legal), Low-resource languages	Optimized for efficiency and specific domain knowledge	0.97	0.69	0.80

Three innovative Language Learning Models (LLMs)-GPT-4, GPT-3.5 and Gemini-were thoroughly tested for their diagnostic capabilities in our exhaustive review. The objective was to evaluate the efficacy of these models in analyzing and diagnosing medical issues using comprehensive symptom descriptions. The results, shown in **Figure 1**, show that the models differ significantly in their diagnostic accuracy and capacities, which helps to explain their possible use in clinical contexts. Based on its remarkable diagnostic accuracy, GPT-4 stood out as the top performer in our investigation. This model has a deep grasp of medical symptomatology thanks to its rigorous training on a mountain of patient data and medical literature. The extensive algorithmic structure and comprehensive data processing capabilities of GPT-4 are demonstrated by its constant and accurate ability to diagnose diseases from symptom descriptions. It is an impressive tool that could change the way doctors diagnose patients and arrange their treatments; it sets a standard for AI-driven medical diagnostics. In terms of performance, GPT-3.5 was right behind, showcasing strong diagnostic abilities. Its ability to transform complicated symptom data into precise health assessments makes it an invaluable tool in medicine, even though it fell short of GPT-4. By offering trustworthy interpretations of medical situations, GPT-3.5 facilitates clinical decision-making and can substantially assist doctors in better understanding and diagnosing patient problems. Its impressive results demonstrate the trustworthiness of LLMs that have undergone thorough training in medical diagnostics and the promise of AI to greatly improve routine healthcare operations.

Comparative Analysis

Table 2 summarizes the performance indicators and **Figure 2** provides a visual explanation of how they are compared. Our investigation sheds light on the diagnostic capacities of GPT-4, GPT-3.5 and Gemini, each of which has its own distinct capabilities. As evidence of its thorough training process covering a variety of medical data, GPT-4's extraordinary number of right answers is noteworthy. The model's success in effectively mapping symptoms to diagnoses and comprehending complicated medical terminology is demonstrated by its superior F1 score, which is a result of its thorough training.

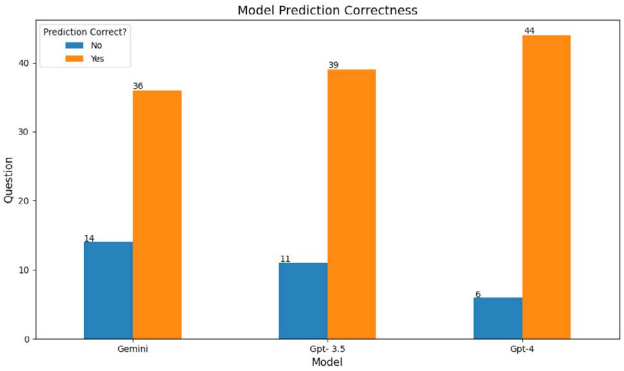
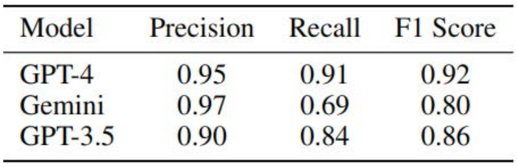


Figure 2: Model Prediction Correctness.

Table 2: Comparative Performance of LLMs in Digital Diagnostics.



Model	Precision	Recall	F1 Score
GPT-4	0.95	0.91	0.92
Gemini	0.97	0.69	0.80
GPT-3.5	0.90	0.84	0.86

Figure 3 presents a visual representation of the models' relative accuracy, drawing attention to GPT-4's superiority in answering questions correctly and demonstrating its remarkable capacity to handle the intricacies of the symptom diagnosis association. By comparing the results visually, we can see that GPT-4 outperforms the other models in terms of accuracy and comprehension. Figure 3 explores the confidence levels linked to each model's predictions, which further improves our analytical viewpoint. Here, the majority of GPT-4 and GPT-3.5's confidence distributions fall into the 'High' group. Gemini stands out for its exceptional accuracy because it tends to provide responses with a high level of confidence, even if it makes fewer predictions overall. Because of the gravity of the consequences for making an incorrect diagnosis, this quality is of the utmost importance in healthcare settings.

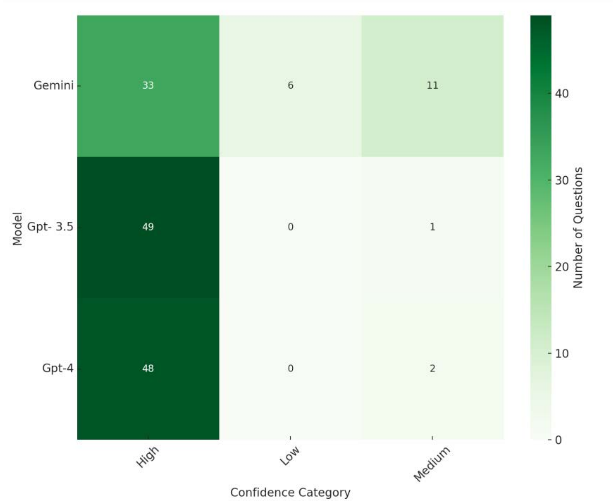


Figure 3: Model Confidence Category.

5. Conclusion

The diagnostic capacities of GPT-4, Gemini and GPT-3.5 are examined in this study, which provides insights, respectively. When it comes to detecting common ailments, the GPT-4 is renowned for its high level of accuracy, which places it at the top of the list of models that were evaluated. Because of its high level of precision, Gemini demonstrates a great deal of promise as a supplementary instrument for digital diagnostics, particularly in activities that call for pinpoint accuracy. Despite being slightly less advanced than other options, GPT-3.5 continues to be a dependable choice. It ranks second in terms of the accuracy of disease prediction. The tremendous advancements in LLM technology and the real benefits it offers to the healthcare industry are highlighted here. The findings presented here shed light on the significant potential that lies within the incorporation of LLMs into digital healthcare systems. In addition to this, they stress the importance of continuous improvement in order to enhance the precision and dependability of these models, which will ensure that they are able to efficiently meet clinical requirements in digital healthcare contexts.

6. References

1. Abbasian M, Azimi I, Rahmani AM and Jain R. Conversational Health Agents: A Personalized LLM-Powered Agent Framework 2024;1.
2. Baharudin N, Mohamed Yassin MS, Daher AM, et al. Prevalence and factors associated with lipid-lowering medications use for primary and secondary prevention of cardiovascular diseases among Malaysians: the REDISCOVER study. In: BMC Public Health 22 (2022) https://doi.org/10.1186/s12889-022-12595-1 .
3. Parker JB and Anderson EE. Patient Data-Sharing for AI: Ethical Challenges, Catholic Solutions. In: The Linacre Quarterly 2020;87(4):471-481.

4. Batsis JA, Mackenzie TA, Emeny RT, Lopez-Jimenez F, and Bartels SJ. “Low Lean Mass With and Without Obesity, and Mortality: Results From the 1999–2004 National Health and Nutrition Examination Survey”. In: The Journals of Gerontology: https://doi.org/10.1093/gerona/glx002 2017;72(10):1445-1451

5. Chiu YY, Sharma A, Lin IW and Althoff T. A Computational Framework for Behavioral Assessment of LLM Therapists 2024

6. Choudhury A and Chaudhry Z. Large Language Models and User Trust: Focus on Healthcare (Preprint). In: Journal of Medical Internet Research http://dx.doi.org/10.2196/56764 2024;26

7. Cui H, Fang X, Xu R, KanX, Joyce CH and Yang C. Multimodal Fusion of EHR in Structures and Semantics: Integrating Clinical Records and Notes with Hypergraph and LLM. 2024

8. J. de Curtò, I. de Zarzà, Roig G, Cano JC, Manzoni P and Calafate CT. LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments. In: Electronics 2023;12(13)

9. Dhakal U, Singh AK, Devkota S, Sapkota Y, Lamichhane B, Paudyal S, et al. GPT-4’s assessment of its performance in a USMLE-based case study 2024;1.

10. Frantzidis CA and Bamidis PD. Description and future trends of ICT solutions offered towards independent living: the case of LLM project. In: Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments. PETRA ’09. Corfu, Greece: Association for Computing Machinery 2009:1-8 https://doi.org/10.1145/1579114.1579173 .

11. Ghosh A, Acharya A, Jain R, Saha S, Chadha A and Sinha S. CLIPSyntel: CLIP and LLM Synergy for Multimodal Question Summarization in Healthcare 2023;38(20)

12. Humphrey BA. Data Privacy vs. Innovation: A Quantitative Analysis of Artificial Intelligence in Healthcare and Its Impact on HIPAA regarding the Privacy and Security of Protected Health Information. PhD dissertation. Robert Morris University, 2021.

13. Yuan J, Tang R, Jiang X, and Hu X. LLM for Patient-Trial Matching: Privacy-Aware Data Augmentation Towards Better Performance and Generalizability. In: American Medical Informatics Association (AMIA) Annual Symposium https://par.nsf.gov/biblio/10448809 .

14. Jin M, Yu Q, Shu D, Zhang C, Fan L, Hua W, et al. Health-LLM: Personalized Retrieval-Augmented Disease Prediction System 2024;1

15. Jo E, Epstein DA, Jung H and Young-Ho Kim. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI ’23. , Hamburg, Germany, : Association for Computing Machinery 2023:1-16.