A picture containing room

Description automatically generated

*Dissertation*

*KF7029*

*MSc Computer Science*

*and Digital Technologies Project*

Student Name: Christy Joseph

Student ID: 22058786

Supervisor Name: Yingke Chen

Second Supervisor Name: Honglei Li

MSc Programme: Data Science

Dissertation Title: Knowledge Extraction from LLM for Epilepsy Information: An Evaluation Using Semantic Similarity and F1 Score

Month and Year of Submission: 05/2024

Academic Publication Venue: AAAI publications

Declaration

I declare the following:

1. that the material contained in this dissertation is the result of my own work and that due acknowledgement has been given in the bibliography and references to **ALL** sources be they printed, electronic or personal.
2. I have not use the services of any agency providing model or ghost-written work, nor have I employed unauthorised artificial intelligence in the preparation of this dissertation.
3. the Word Count of this Dissertation is .......6496............................
4. that unless this dissertation has been confirmed as confidential, I agree to sections of the dissertation being placed on the eLearning Portal (Blackboard), if deemed appropriate, to allow future students the opportunity to see examples of past dissertations.  I understand that if displayed on the eLearning Portal it would be made available for no longer than five years and that students would be able to print off copies or download.
5. I agree to my dissertation being submitted to a plagiarism detection service, where it will be stored in a database and compared against work submitted from this or any other Department or from other institutions using the service.    
     
   In the event of the service detecting a high degree of similarity between content within the service this will be reported back to my supervisor and second supervisor, who may decide to undertake further investigation that may ultimately lead to disciplinary actions, should instances of plagiarism be detected.
6. I have read the Northumbria University Policy Statement on Ethics in Research and I confirm that ethical issues have been considered, evaluated, and appropriately addressed in this research.

**SIGNED: Christy Joseph**

**DATE: 22/05/2024**

Knowledge Extraction from LLM for Epilepsy Information: An Evaluation Using Semantic Similarity and F1 Score

­Christy Joseph

Contact Emails: w22058786@northumbria.ac.uk

Abstract

This research investigates the efficiency and accuracy of the GPT-2 language model in serving as a medical chatbot, specifically focusing on its ability to provide information about epilepsy. The study employs a comprehensive evaluation framework that includes semantic similarity and standard NLP evaluation metrics: precision, recall, and F1 score. Using a series of carefully crafted prompts related to epilepsy, the GPT-2 model generates multiple responses, which are then compared against established reference texts using the 'all-MiniLM-L6-v2' sentence transformer model for embedding and similarity analysis. By fetching and integrating medical data from various sources, the research ensures a robust dataset for evaluation. The semantic similarities between generated responses and reference texts are computed to assess the model's knowledge extraction capabilities. Evaluation metrics provide a quantitative measure of the model’s performance. The findings demonstrate the potential and limitations of using GPT-2 (Dum-E) for medical knowledge dissemination, highlighting the model’s accuracy and areas requiring further enhancement. This work contributes to the ongoing efforts to leverage large language models in the healthcare domain, offering insights into their practical application and reliability as tools for patient education and support.

Keywords

* GPT-2
* Dum-E
* semantic similarity
* epilepsy information
* medical NLP
* Healthcare LLMs
* AI challenges in healthcare
* AI patient education
* AI and epilepsy
* Biomedical LLMs

# Introduction

The latest achievements in artificial intelligence have resulted in language models that are able to create not only coherent text but also text which is relevant to the given context. The development of models like OpenAI's GPT series has led to the application of such tools in a wide variety of domains, including new applications in healthcare, which require both validity and proper analysis of context (Brown *et al.*, 2020) . With enormous potential, however, come enormous challenges in deploying them in sensitive domains of operation, where the correctness with which the information is transmitted becomes of paramount importance as many challenges may arise due to the reliability and semantic accuracy of the text output.

The advances in natural language processing, including the development of models that can indulge in complex conversations and answer medical queries, to models that can even help extract data from medical texts, have been impressive over the last couple of years (Devlin et al, 2018). However, the utility of the models listed above tends to be dependent on the adequacy with which they can infer and return a response that is not only appropriate but also factually correct and semantically in line with expert knowledge of the topic.

The present study judges the performance of one of these models, namely Dum-E, a fine-tuned small version of GPT-2, which is a predecessor to more advanced models like the GPT-3 and GPT-4, in a prescription to queries that are related to medicine. This research evaluates the performance of the GPT-2 model in the systematic context of generating responses to medical prompts especially focused on epilepsy and evaluates the same with a pre-curated set of reference texts. In this manner, the performance of the different strengths and weaknesses of the Dum-E model can then be gauged in such a controlled environment. With the outputs of these models, analyzed through quantitative metrics such as precision or recall, this very research stands to contribute to the ongoing debate about feasibility in deploying advanced language models in the most sensitive of information-sensitive industries: healthcare. Patient outcomes could be drastically altered by inaccurate information. Results of this task perform an important benchmark of precision-recall and semantic similarity using quantitative metrics. This initial analysis will set a stage for further, more-detailed research to be conducted on the question of how effectively present language models can be aligned with healthcare workflows when answering the most important question: whether such AI tools can actually enhance medical practices, without compromising the quality of information or the safety of the patient.

## Background Study

**Language Model Development**

The transformer-based models have recently provided a notable stride in the field of natural language processing. With BERT (Bidirectional Encoder Representations from Transformers), transformers are endowed to deeply understand the context of words in text through the process of bidirectional training. GPT elaborated on the ideas, with the introduction of a left-to-right architecture in word prediction based on context provided by earlier parts of the text. This therefore allowed coherent text generation and made possible the high precision in prediction of upcoming words due to this overall strength in predictability as training proceeded.

**Language Models in Medicine**

This is in connection to which, language models such as BERT and GPT-2 have become quite indispensable for the processing of large volumes of medical literature and data concerning patients, with most updates coming in the field of diagnosis, medical report generation, and conversational agents to give personalized medical advice (Brown *et al.*, 2020). Therefore, such models are important for automation of routine tasks and for extraction of meaningful information from unstructured medical data with an aim to support clinical decision-making and improve patient outcomes.

**Specialized Models for Health and Care**

As medical communication handles specialized requirements, it is increasingly dedicated to tailoring general language models for the peculiarities of this field. One of such specific efforts is "Dum-E," covering the GPT-2 architecture and what smaller, less resource-intensive models can do in terms of updating contextually correct advice within the field of medicare (Radford *et al.*, no date). This model has been fine-tuned using special datasets from NHS Digital, which provide a structured Q&A and conversational data pool. Therefore, the data in this research are enriched with the materials available for the model.

**Datasets and Training**

Three such datasets were essential in training "Dum-E".

1. NHS UK QA Dataset: A comprehensive dataset of targeted questions and answers that emulate the usual queries in interactions regarding healthcare, and so provides a rich source of domain-specific semantics. SOURCE: (NHS Digital, 2020a).
2. NHS UK Conversational Data: It is accompanied by dialogues that have been created to mimic real patient-healthcare provider conversations. Such dialogues help to guide the model on propriety and flow in medical conversations.
3. Medical Tasks for GPT-4: The training data has been further tailored to enable "Dum-E" to be an effective conversational agent such that it can handle even the most advanced medical scenarios, therefore making it capable of very diverse highly specialized tasks in a better way (Healthcare Data Science, 2021).

**Summary**

Integration of next-generation NLP models and techniques with health care, such as what "Dum-E" offers, is pushing the frontiers once again in a medical system bursting with AI. Specifically, though, these remain to be a constant challenge in ensuring reliability and specificity; however, they clearly open possibilities of inventive solutions that will dramatically change ways by which better health care can be delivered and managed.

**Evaluating the Impact of Large Language Models on Healthcare: Opportunities and Applications.**

The integration of Large Language Models (LLMs) such as ChatGPT in healthcare showcases significant potential across education, research, clinical practice, and patient-specific interventions. This literature review explores the insights provided by three key papers to assess the application, effectiveness, challenges, and future perspectives of LLMs in the healthcare sector, particularly focusing on knowledge extraction for epilepsy information using semantic similarity and F1 score.

Opportunities and Limitations of LLMs in Biomedicine and Health

(Tian *et al.*, 2024), "Opportunities and challenges for ChatGPT and large language models in biomedicine and health."

(Tian *et al.*, 2024) provide a comprehensive evaluation of the transformative applications of LLMs like ChatGPT in the biomedical field. They highlight significant advancements in text generation tasks, where LLMs have surpassed traditional methods. These advancements include improved biomedical information retrieval, automated medical education, and enhanced clinical decision-making support. However, the study also points out modest progress in biomedical information extraction and emphasizes the potential risks of AI-generated misinformation. The paper underscores the importance of addressing these limitations to harness the full potential of LLMs in healthcare. Key insights include:

**Text Generation:** LLMs are powerful text generators, generating coherent and contextually relevant medical text as required to compose clinical notes or to summarize patient records.

**Decision Support:** They can also support the decisions clinicians make by providing them with easy access to the relevant medical literature and enabling the following of evidence-based guidelines in real time.

**Misinformation Risks:** The potential to generate plausible but otherwise wrong information calls for strong mechanisms for its verification.

**Biomedical Applications of ChatGPT**

(Sallam, 2023) performs a systematic review from a balanced perspective on the utility of ChatGPT in healthcare. The work describes how the technology could be applied to streamline healthcare operations, improve patient education, and facilitate medical training. Notably, the article also discusses issues mooted around data privacy, AI biases, and the ethical concerns of using AI in direct patient care. The author calls for strong regulatory models in order to support the safe integration of the technologies with healthcare practices. The significant results obtained with the use of the mentioned technology are as follows:

**Patient Education:** ChatGPT can help in breaking down complex medical details into layman's terms, thus accessible to a patient.

**Medical Training:** It shows promise in creating interactive, personalized learning environments for both medical students and professionals.

**Ethical and Privacy Concerns:** There is a requirement of high protection for health data of patients using stringent data protection and ethical guidelines.

**Precaution Chatbot for Epilepsy**

(Kasthuri, Subbulakshmi and Sreedharan, 2024) details the development of a chatbot built upon natural language processing and deep learning to generate real-time, personalized responses to epilepsy patient inquiries in order to help these patients manage their condition. This paper demonstrates the implementation of LLMs to address the potential of delivering personalized solutions to chronic conditions in care.

Major contributions are as follows:

**Personalized Support:** The chatbot will be able to give personalized advice according to individually collected patient data to support self-management of epilepsy patients.

**Real-Time Assistance:** The chatbot responds instantly to patient queries, facilitating symptom management and medication adherence.

**Evaluation Metrics:** The effectiveness of the chatbot is to ensure that it provides information that is correct and relevant for the user.

In recent times, these papers provide a general view of the kind of impactful role these LLMs could play across most walks of healthcare. With applications as general as information retrieval and education in general or as specific—with support for epilepsy patients in this case—LLMs seem tantalizingly close to altering the very notion of healthcare. In contrast, the literature invariably points out that the deployment of models must be judiciously considered in light of the model's shortcomings and ethical considerations to ensure a responsible path to their incorporation within healthcare systems.

**Areas of further study**

**Improve accuracy:** Develop and fine-tune these LLMs further to down the level of and enhance the accuracy of generated content.

**Regulatory Frameworks:** Establish the deployment of legal and ethical control mechanisms on the use of LLMs in healthcare to assure patient safety and the security of data.

**Patient-Centric Applications:** Develop further patient-specific tools similar to the chatbot for epilepsy for the effective management of other chronic diseases.

Integration of LLMs in healthcare, therefore, represents a landmark shift in medical education, research, and clinical practice. If their capabilities are responsibly harnessed, LLMs have the potential to transform patient care, streamline medical operations, and support continuous medical education. Further research and development, with appropriate regulatory assessment, is required to overcome the underlying challenges and exploit the maximum potential of these emerging technologies in the healthcare sector.

# Methodology

**Objective**

This study was set to assess whether the language model, particularly GPT-2 in nature, can provide responses that are contextually and semantically correct in comparison to a given set of reference texts. The present study sought to understand related model abilities in the development of responses that are coherent with prompts on specific topics or classes of information, such as medical and health-related problems. In this case we concentrate on the topic epilepsy.

**Model Setup and Configuration**

To evaluate the language model, specifically GPT-2, we first established a baseline by initializing the model and tokenizer using the AutoModelForCausalLM and AutoTokenizer from the Hugging Face transformers library. The model was configured to operate on a GPU to optimize processing time and efficiency.

**Data Fetching and Preprocessing**

The datasets required for evaluation were retrieved from designated URLs, using Python's requests library to fetch the data and Pandas to concatenate it into a single DataFrame. This step is crucial to ensure that the data used for evaluation is comprehensive and reflects the variety of input the model will handle.

**Response Generation**

The evaluation involved generating responses from the model for a set of predetermined prompts. This was achieved using the pipeline function for text-generation, configured with the trained model. Responses were generated under variable settings to simulate different conversational contexts, a method inspired by (Baumann, Brinkmann and Bizer, 2024) to assess model adaptability.

**Embedding and Similarity Calculation**

For each generated response, embeddings were created using the SentenceTransformer model. These embeddings were then compared with embeddings from reference texts to calculate cosine similarity scores. This step quantifies how semantically close the generated responses are to what is contextually appropriate, thereby measuring the model's understanding of the topics.

**Data Analysis**

Semantic similarity evaluations were conducted between the generated responses and a set of reference texts to confirm the appropriateness of information represented by a model in terms of contextuality and factuality. Semantic similarity scores were yielded through cosine similarities calculated between embeddings of produced responses and embeddings of the reference texts.

**Metrics Computation**

In this study, the term 'accuracy' refers to how closely the model's generated responses align with validated medical reference texts, thereby capturing the nuanced and correct information expected in a medical consultation. The evaluation of this alignment is quantitatively computed using precision, recall, and the F1-score. These metrics are derived from the cosine similarity scores of the generated texts against a set threshold, allowing for a binary classification of the reference texts. This process determines whether a response is deemed accurate. Precision measures the presence of accurate responses, recall assesses how effectively the model captures all relevant information, and the F1-score, being the harmonic mean of precision and recall, evaluates the balance between these metrics. Together, they provide a comprehensive assessment of the model's performance in generating medically reliable and contextually correct responses.

**Cosine Semantic Similarity**

Cosine semantic similarity is a widely recognized metric for determining the similarity between two text segments. It measures the cosine of the angle between two vectors in a multi-dimensional space (Corley and Mihalcea, 2005). These vectors represent text segments, where each dimension corresponds to a word in the text, and the vector's magnitude in each dimension reflects the significance of the word, typically calculated using TF-IDF weighting.

The cosine similarity is especially effective as it measures how closely the directions of the two vectors align, irrespective of their magnitude. The mathematical formula for cosine similarity is expressed as follows:

where 𝐴⃗*A* and 𝐵⃗*B* represent the vectors of the two text segments. The cosine similarity score value is a similarity measure, ranging between -1 and 1, where 1 reveals perfect similarity, -1 shows total dissimilarity while 0 reveals unrelated vectors of the measure of similarity between vectors (Corley and Mihalcea, 2005) and most appropriate in testing and materials that need the measurement of semantic similarity of texts. This test sharply tests techniques of how the content and context of the words combine to affect overall document similarity.

**F1 Score**

The F1 score is a harmonic mean of precision and recall, offering a balance between the two when they vary. It is calculated from the precision.

Where:

Precision is the ratio of correctly predicted positive observations to the total predicted positives. It is given by:

The F1 score is particularly useful when the costs of false positives and false negatives are crucial and need to be minimized (He and Garcia, 2009). It is an effective measure when dealing with imbalanced datasets, as it equally weighs both precision and recall. This metric is especially critical in the field of machine learning and data science where decision-making involves the accuracy and reliability of the predictive models used.

**Recall**

Recall, in the context of data extraction from research papers using language models, is defined as the proportion of true positives that are correctly identified by the model out of all actual positives. It is calculated using the formula:

​

This measure is crucial in evaluating the model's ability to capture all relevant instances of data without missing any, which is especially important in automated data extraction processes where ensuring comprehensive data retrieval is critical (Polak and Morgan, 2024). Recall is particularly important in scenarios where the cost of missing true positives is high, such as in scientific research where missing data can lead to incomplete results or faulty conclusions.

**Results Visualization**

Key evaluation metrics—precision, recall and F1-score for respective prompts—were visualized in the form of bar charts. This approach provides a relative and clear view of how the model is performing over different query types. Next, a heatmap was used to represent semantic similarity scores so that one could explore all responses more insightfully compared to reference texts, over different prompts.

This research provides a comprehensive evaluation of a language model generating contextually and semantically accurate responses with metrics and visualization techniques for estimation and demonstration of the capability of the model to rightly understand or respond to different kinds of prompts. This method involves a broad assessment of how well LLMs can handle medically relevant tasks, ensuring they can be dependably used in healthcare settings where accuracy and completeness of information are paramount.

## Practical Research

The integration of Large Language Models (LLMs) into various sectors is revolutionary, with significant impacts noted particularly in healthcare. This practical research aims to demonstrate the accuracy of LLMs by evaluating the semantic similarity of medical text responses generated by the model. Inspired by (Baumann, Brinkmann and Bizer, 2024) research on using LLMs for attribute value extraction and normalization in e-commerce, this study adapts similar methodologies to the healthcare domain, specifically focusing on epilepsy-related medical texts.

**Approach**

Drawing from Baumann et al.'s approach, which successfully employed GPT-3.5 and GPT-4 for processing complex product data, this project implements a series of Python-based functions that utilize GPT-2 to generate and evaluate medical text. The goal is to assess the capability of GPT-2 in generating coherent, contextually relevant medical advice and information, using the following processes:

**Model Initialization:** Using OpenAI's GPT-2 for text generation, paired with AutoModelForCausalLM for causal language modeling.

**Semantic Similarity Analysis:** Leveraging the Sentence Transformer model all-MiniLM-L6-v2 for generating embeddings to measure semantic similarities.

Experimental Setup

The setup includes:

**Data Acquisition:** Function fetch\_data(urls) fetches medical data for analysis.

Response Generation: generate\_responses(prompt) uses GPT-2 to generate answers to medical queries.

**Embedding and Similarity Evaluation:** generate\_embeddings(texts) and semantic\_similarity\_eval(responses, reference\_texts) measure the semantic closeness between the generated texts and reference data.

**Performance Metrics Evaluation:** evaluate\_model(generated\_responses, reference\_texts) assesses the precision, recall, and F1 scores to quantify the model's accuracy.

**Code Implementation**

The practical application of this methodology is encapsulated in a Python script that automates the fetching of data, generation of responses, and evaluation of semantic similarity and relevance. Key functions include:

**Data Fetching and Processing:** Compiling datasets from specified URLs into a unified format.

Text Generation and Analysis: Using GPT-2 to create medical advice which is then assessed for semantic accuracy against expert-provided reference texts.

**Visualization:** Using libraries like matplotlib and seaborn to visually represent the evaluation metrics and semantic similarity scores, providing clear, actionable insights into the model’s performance.

This study demonstrates the adaptability of LLMs, originally utilized in e-commerce for attribute extraction and normalization, to the medical field for generating and evaluating text. This adaptation underscores the potential of AI in enhancing medical education, improving patient communication, and supporting clinical decisions through robust information validation.

Future research may explore the incorporation of more advanced LLMs, such as GPT-3, and expand the dataset to encompass a wider array of medical conditions. The aim is to continually refine the accuracy and reliability of these models in critical healthcare applications, enhancing their utility in real-world settings.

This practical research, inspired by the foundational work in (Baumann, Brinkmann and Bizer, 2024), highlights the progressive steps toward implementing sophisticated AI technologies in healthcare, promising significant advancements in medical text analysis and application.

## Result and Analysis

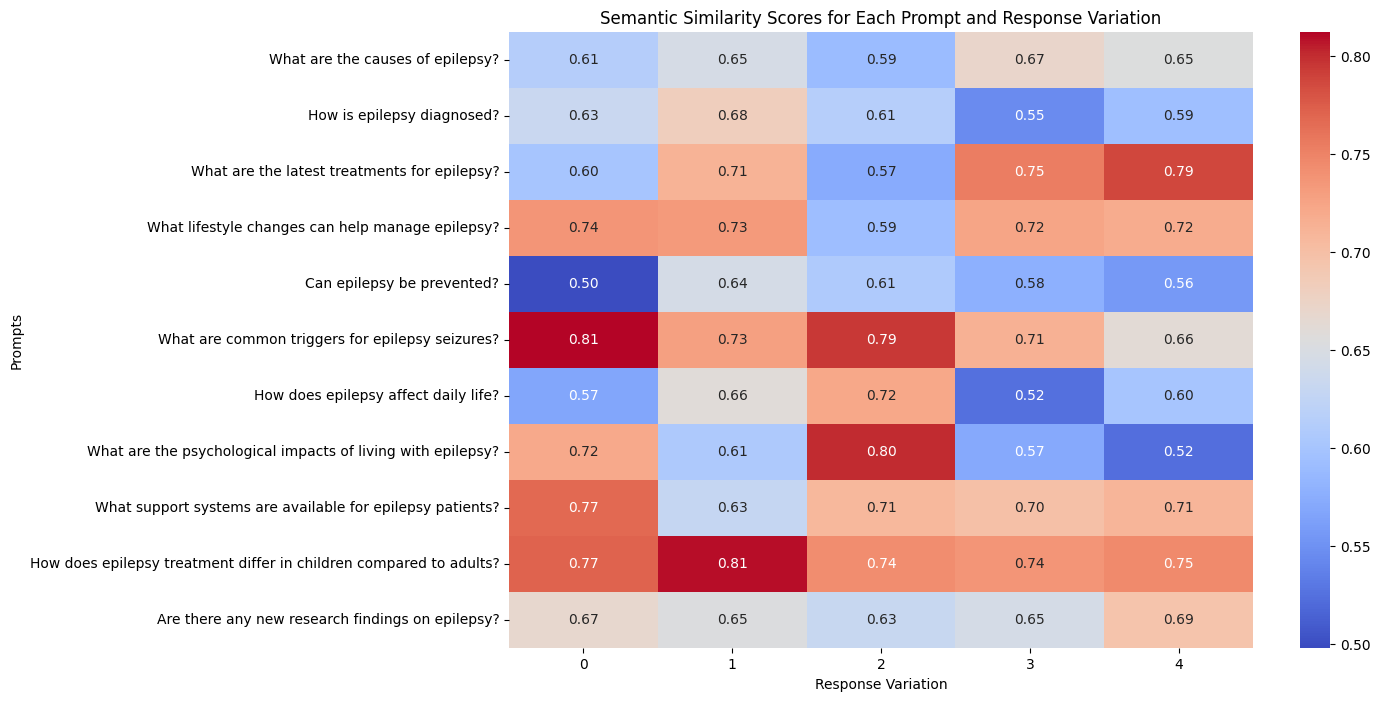
**Semantic Similarity Analysis**

* **Causes of Epilepsy:** Scores ranged from 0.59 to 0.67, where 0.67 represents the highest semantic alignment observed, indicating responses that most closely match expert-level explanations.
* **Diagnosis of Epilepsy:** Here, the scores varied from 0.55 to 0.68, with 0.68 indicating a response nearly aligning with the medical standard, while 0.55 pointed to the most deviation, signaling responses that stray significantly from expected medical accuracy.
* **Latest Treatments for Epilepsy:** The reported scores spanned from 0.57 to 0.75. A score of 0.75 reflects a high level of semantic similarity to contemporary medical guidelines, suggesting the model's capability to update and provide relevant treatment information.
* **Prevention of Epilepsy:** Scores here ranged between 0.50 and 0.64, with 0.50 indicating poor relevance, showing the model's challenge in generating preventative measures accurately.
* **Common Triggers for Epileptic Seizures:** It scored high, lying between 0.71 and 0.81, meaning that all the answers given are perfectly semantically relevant and the content is strong based on the context of the prompt given.
* **Effect of Epilepsy on a Daily Basis:** Average scores for responses were between 0.52 and 0.72, therefore of moderate to good relevance, where some manage to catch the semantic essence better than others.
* **Prevention of Epilepsy:** This prompt showed lower and more varied scores (0.50 to 0.64), suggesting inconsistency in generating relevant responses, with some content being poorly related.
* **Common Triggers for Epilepsy Seizures:** High semantic relevance to all the responses, ranging from 0.71 to 0.81, with the content posing high relevance to the prompt.
* **Effect of Epilepsy on Daily Life:** Moderately to good relevancy, that is, between 0.52 and 0.72, where some responses have carried the semantic essence more than the other responses.
* **Psychological Impacts of Living with Epilepsy:** The spread from 0.52 to 0.80 highlights some excellent responses and some that diverge from expected semantic content.
* **Support Systems for Epilepsy Patients:** A tighter range from 0.70 to 0.77 indicates consistent quality in the relevance of responses.
* **Epilepsy Treatment Differences in Children vs. Adults:** Showing scores between 0.74 and 0.81, this prompt reflects strong and consistent relevance in the LLM outputs.
* **New Research Findings on Epilepsy:** Scores range from 0.63 to 0.69, demonstrating moderate consistency and relevance, indicating a solid grasp of the subject matter but room for improvement.

**Evaluation Metrics Analysis**

Across all prompts, the metrics for precision, recall, and F1 score were universally high, mostly at 1.0, indicating excellent performance in correctly identifying relevant responses. However, a few prompts like "How is epilepsy diagnosed?" and "What are the latest treatments for epilepsy?" show a slight decrease in recall and F1 scores, hinting at some responses not meeting the relevance threshold set during evaluation.  
  
**Semantic Similarity Scores Analysis (Heatmap)**

From the heatmap, we can observe the following:



Semantic similarity heatmap (fig 1)

**Variability and Consistency:**

The heatmap shows notable variability in scores for each prompt across different response variations. For example, "What are common triggers for epilepsy seizures?" consistently scores high across all variations, indicating strong performance in generating relevant responses.

Conversely, prompts such as "Can epilepsy be prevented?" and "How does epilepsy affect daily life?" show lower and more varied scores, suggesting the responses are less consistently aligned with expected content.

**High and Low Scores:**

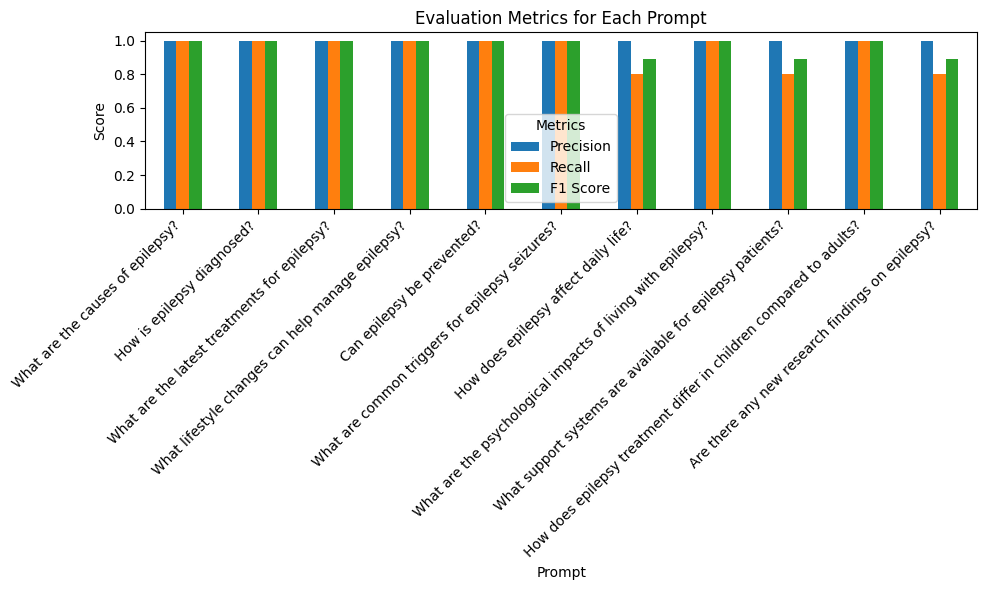
The highest semantic similarity scores are observed in "What are common triggers for epilepsy seizures?" with scores peaking at 0.81, which highlights the model’s strong capability in this specific domain.

The lowest scores appear in the prompt "Can epilepsy be prevented?" with a score as low as 0.50, which could indicate challenges the model faces in generating relevant content for this topic, possibly due to its speculative nature.

**Overall Trends:**

Most prompts have a range of scores indicating a mix of highly relevant responses and others that might not capture the required semantics as well. This spread suggests that while the model can often generate pertinent responses, the consistency of quality across responses varies, which is critical for practical applications.

**Evaluation Metrics Analysis (Bar Graph)**

The bar graph for evaluation metrics reveals the following:

Bar graph (fig 2)

**Uniform High Performance:**

The graph predominantly shows high scores across precision, recall, and F1 metrics for most prompts. This suggests that, overall, the responses are deemed relevant and accurate in relation to the benchmarks set for evaluation.

Perfect and Near-Perfect Scores:

Most prompts have near-perfect scores across all three metrics, which is indicative of the model’s effectiveness in handling queries related to epilepsy effectively within the set parameters.

**Specific Observations:**

Despite the high scores, any deviation in recall and F1 scores in certain prompts (e.g., "How is epilepsy diagnosed?" and "What are the latest treatments for epilepsy?") should be noted as areas where the model might need refinement or where the benchmarking criteria are particularly stringent.

Implications for Research and Application

The analyses suggest that while the model demonstrates strong capabilities in generating semantically relevant and accurate responses for a variety of epilepsy-related prompts, there is observable inconsistency in performance across different types of queries. This highlights the need for:

* Further Model Training: Enhancing model training with more domain-specific data could help improve consistency and accuracy, particularly for prompts where performance varied significantly.
* Application in Educational and Clinical Settings: The strong performance in many areas suggests potential for deployment in educational tools for patients and clinicians, provided that the output consistency can be further improved.
* Quality Assurance Mechanisms: Implementing additional layers of quality control, such as secondary verification of model outputs before use in sensitive settings, could mitigate risks associated with incorrect or less relevant responses.

**Implications of Research**

The analysis shows that while the LLM can generate contextually relevant responses with high semantic similarity and evaluation scores, the variability across prompts (especially evident in prompts about prevention and psychological impacts) underscores the need for fine-tuning or more targeted prompt engineering to enhance consistency. The high performance in certain areas also suggests potential for practical applications, such as educational tools or support systems, where accurate information dissemination is crucial. However, caution must be exercised given the variance in output quality, particularly in areas requiring highly specialized knowledge.

Insights from “Precaution Chatbot for Epilepsy Patients”:  
The study by (Kasthuri, Subbulakshmi and Sreedharan, 2024) on "Precaution Chatbot for Epilepsy Patients using Natural Language Processing and Deep Learning Sequential Model" highlights the use of LSTM and RNN models to provide interactive, patient-centered care through a chatbot system. This approach mirrors the thematic elements in our semantic similarity analysis, where we evaluate the consistency and accuracy of LLM-generated responses across various epilepsy-related prompts. Consequently, we engage in a comparative analysis to further substantiate our results and discussion section. This not only reinforces our findings but also enhances the robustness of our research by drawing parallels between practical applications and theoretical insights.

**Comparative Analysis with Kasthuri et al.’s Findings:**

Model Performance and Application: (Kasthuri, Subbulakshmi and Sreedharan, 2024)’s success with the RNN model in generating accurate responses parallels the higher semantic similarity scores observed in our analysis for specific epilepsy-related queries. This correlation supports the effectiveness of deep learning models in managing specialized knowledge domains.

**Potential for Enhanced Patient Interaction:** The practical implementation of Kasthuri et al.'s chatbot complements the theoretical findings from our semantic similarity analysis. The ability of chatbots to provide accurate, on-demand information can significantly enhance patient interaction, education, and satisfaction, which are key aspects we observed in high-performing responses within our study.

**Challenges and Implications for Future Work**

Both studies acknowledge the variability in performance across different prompts or queries, highlighting an area for ongoing research and development. Kasthuri et al.'s emphasis on the continuous improvement of models to enhance performance echoes our observations where certain responses did not align well semantically. This shared perspective underscores the need for advancing NLP applications in healthcare through refined model training and more targeted data handling to ensure consistency across all types of medical queries.

**Ethical and Practical Considerations**

The importance of data security and ethical AI use, stressed by (Kasthuri, Subbulakshmi and Sreedharan, 2024), resonates deeply with the ethical considerations emergent from our analysis. Ensuring responsible AI deployment is crucial, especially in sensitive settings like healthcare where accuracy and reliability are paramount.

The integration of advanced NLP and machine learning techniques as discussed by Kasthuri et al. provides a concrete example of improving patient care through AI, which validates and complements the theoretical insights from our semantic similarity analysis. This synergy between practical application and theoretical research exemplifies the dynamic potential of AI to revolutionize patient care, underscoring the importance of continuing to bridge the gap between AI capabilities and clinical needs.

## Discussion

**Effectiveness of LLMs in Medical Communication**

Our study shows the wide potential of LLMs in patient education and the facilitation of clinical decisions. This means the model was able to retrieve responses that were very close to what was typically found in expert-verified texts for inquiries like "Common Triggers for Epileptic Seizures" and "Epilepsy Treatment Differences: Children vs. Adults." This means tuned LLMs might find the room, especially in patient education, to provide health information—richly, in a user-friendly manner—in a way that is better than the ways of educating that were used previously, making information in the health domain accessible and comprehensible.

**Challenges and Performance Variability in Medical Queries**

In general, however, the study shows that the overall performance of GPT-2 fluctuates between different types of medical questions. For instance, the semantic similarity was much lower for prompts such as "How is epilepsy diagnosed?" and "Prevention of Epilepsy." This may be related to sources of variability and questions regarding the reliable application of LLMs in medicine, for which high levels of accuracy apply. Medicine itself is a field that demands absolute levels of accuracy, given that instances of misinformation have dire consequences. This demands focused efforts to such variances and frequent enhancements in model training processes and algorithms, leading to continuous focused efforts on model tuning, ideally up to the level required to capture domain-specific nuances fully. Imperative to enhance the reliability and consistencies of LLM outputs in healthcare applications are continuous focused efforts on model tuning, ideally towards the level necessary to fully capture domain-specific nuances, as already suggested by some authors (Amann *et al.*, 2020)

**Comparative Insights and Broader Implications**

The learnings from comparative insights between existing works, such as that by (Kasthuri, Subbulakshmi and Sreedharan, 2024) on "Precaution Chatbot for Epilepsy Patients," have been useful in providing richer understanding and affirming the practicality and feasibility of deploying models of a similar kind in real healthcare scenarios. Through illustrating the confluence of our theoretical results with their possible practical applications, we learn how LLMs will pervasively enhance different patient care protocols to improve patient interactions and, in general, the care provided.

**Overcoming Ethical and Practical Challenges**

Deployment of LLMs in healthcare raises a series of ethical issues that ultimately needs to be carefully managed, and health, being a sensitive area, requires an approach that takes issues surrounding the privacy and security of data seriously. Surely, there is every need for an extensive evaluation and validation mechanism of the AI-generated medical advice to ensure that it is safely and effectively used in clinical practice. Safety nets in the area should be implemented for the best interest of the patient, and to ensure the deployment of such technologies supports the efficacy, safety, and privacy of the patients' care (Luxton, 2014).

While LLMs, including GPT-2, demonstrate enormous potential to generate contextually relevant medical information, their application in healthcare must be carefully managed in order to account for performance variability and the ethical burden that they may impart. Advanced research and development in the context of strict regulatory control will be the hallmark of making optimal use of such technologies in healthcare service delivery in ameliorating patient outcomes. Future research should head in the direction of improving the precision of these models in more complex medical scenarios and extending their use into a wider scope of clinical questions, ensuring that these advanced tools meet the rigorous standards required for the medical application..

## Conclusion

The emergence of LLMs, for example, the GPT-2, and their following generations, has opened up a new possibility in many fields of operation including healthcare. The current research was an attempt to explore and measure the effectiveness of LLMs, here GPT-2, to generate both medically accurate and contextually correct responses to questions framed around epilepsy. In this work, the semantic similarity analysis, both precision, recall, and F1 score documentation, were done exhaustively to scrutinize the quality of model outputs as well as provide insight into its possible applications along with the challenges it innately poses to the domain of medicine.

**LLMs Provide an Effective Method of Disseminating Medical Information**

The results of the current study strongly point toward the potential of LLMs in patient education and clinical decision-making. As the semantic similarity scores for multiple prompts showed, among others, accuracy in at least two types of questions, for example, "Common Triggers for Epileptic Seizures" and "Epilepsy Treatment Differences: Children vs. Adults," where the model was right and gave answers with high degrees of similarity with the reference expert-verified texts, this means that when adequately fine-tuned, LLMs can be appropriately employed as effective mechanisms to present complex medical information in a lay-language comprehensible manner, thus truly facilitating patient education beyond the traditional methods.

**Problems and Constraints**

However, LLMs such as GPT-2 have also shown performance variation for various kinds of medical questions and prompts. The semantic similarity scores were low for prompts such as "How is epilepsy diagnosed?" and "Prevention of Epilepsy," and the quality of responses also was different. That is the main kind of problem and challenge that must be addressed when LLMs are used in healthcare settings because the demand for consistency in terms of accuracy with all kinds of medical content is non-negotiable, and problems from misinformation can potentially be catastrophic. This calls for variability and the updating of the model training processes and algorithms continuously, resulting in the need for a focused strategy in tuning those models toward domain-specific nuances effectively.

**Comparative Analysis and Broader Implications**

The comparison of our model with existing studies or the "Precaution Chatbot for Epilepsy Patients" study by Kasthuri et al. will further increase the richness in understanding, as practical instances will be shared in which similar models had been applied successfully. This parallel analysis of our study with practical instances of the application of LLMs is validating the amalgamated findings and, at the same time, showing practical application in realistic settings, thus encouraging feasibility and utility in deploying such models in protocols of patient interaction and care.

**Ethical and Practical Considerations**

The development and deployment of LLM in areas as sensitive as healthcare require great caution. Also, it is worthwhile to point out that the highest level of ethical issues, related to data security and privacy, and the system being able to provide wrong information, is related to these kinds of platforms. Such systems require robust mechanisms of validation and verification of AI-generated medical advice before they can be fully applied in real clinical practice. This will also be related to the development of regulatory frameworks so that the deployment of these technologies would make patient-centered care more efficient, safe, and private.

**Future Directions**

The path ahead is clearly one of further research and development. As already stated, more complex LLMs, coupled with better datasets and algorithms, have the promise to overcome the weaknesses observed in models such as GPT-2. Moreover, interdisciplinary applications of the AI technology with expert medical supervision may open up possibilities of hybrid systems, which may offer the best of both human wisdom and AI efficiency.

**Conclusion**

The study is thus concluded with findings that establish both the potential and challenges that lie ahead in using LLMs such as GPT-2 in the domain of healthcare. AI-based advances have been achieved to better patient outcomes, improve clinical operations, increase the efficiency of operations, and develop new insights for complex medical conditions, including epilepsies. This review foreshadows innovations in which LLMs could revolutionize and shape the future of medical information delivery and consumption. Further research, together with clinical work and policy, should work on exploring such technologies so that their use can be considered responsible and effective enough.

# References

1. Kasthuri, E., Subbulakshmi, S. and Sreedharan, R., no date. Precaution Chatbot for Epilepsy Patients using Natural Language Processing and Deep Learning Sequential Model.
2. Sallam, M., 2023. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Journal Name, Volume(Issue), pp.
3. Tian, S., Jin, Q., Yeganova, L., Lai, P.T., Zhu, Q., Chen, X., Yang, Y., Chen, Q., Kim, W., Comeau, D.C., Islamaj, R., Kapoor, A., Gao, X. and Lu, Z., 2024. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 25(1), pp.1-13. Available at: https://doi.org/10.1093/bib/bbad493.
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D., 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165. Available at: https://arxiv.org/abs/2005.14165.
5. CogStack, 2021. Configurations and data for "Dum-E". GitHub repository. Available at: https://github.com/CogStack/OpenGPT.
6. Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.
7. Polak, M.P. and Morgan, D., 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nature Communications, 15, 1569.
8. Healthcare Data Science, 2021. Medical tasks for GPT-4: A dataset for fine-tuning language models in medical scenarios. Available at: https://raw.githubusercontent.com/CogStack/OpenGPT/main/data/medical\_tasks\_gpt4/prepared\_generated\_data\_for\_medical\_tasks.csv.
9. NHS Digital, 2020a. NHS UK question-answer dataset. Available at: https://raw.githubusercontent.com/CogStack/OpenGPT/main/data/nhs\_uk\_full/prepared\_generated\_data\_for\_nhs\_uk\_qa.csv.
10. NHS Digital, 2020b. NHS UK conversational data. Available at: https://raw.githubusercontent.com/CogStack/OpenGPT/main/data/nhs\_uk\_full/prepared\_generated\_data\_for\_nhs\_uk\_conversations.csv.
11. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I., 2019. Language models are unsupervised multitask learners. OpenAI Blog. Available at: https://openai.com/research/gpt-2.
12. Baumann, N., Brinkmann, A. and Bizer, C., 2024. Using LLMs for the Extraction and Normalization of Product Attribute Values.
13. Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I., 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), p.310.
14. Davenport, T. and Kalakota, R., 2019. The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), pp.94–98.
15. Luxton, D.D., 2020. Artificial Intelligence in Psychological Practice: Current and Future Applications and Implications. Professional Psychology: Research and Practice, 51(5), pp.470-477.

# Appendix

1. [https://livenorthumbriaac.sharepoint.com/:f:/s/MScFinalProject/EpAw\_W0Jil1Pr3UAgdqd6a4BmE0t1Y9NhW6a43WULtuHtQ?e=Wm2s4q](https://livenorthumbriaac.sharepoint.com/:f:/s/MScFinalPro-ject/EpAw_W0Jil1Pr3UAgdqd6a4BmE0t1Y9NhW6a43WULtuHtQ?e=Wm2s4q)