Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 22 (2024) 389-402

Original

Development of a ChatBot model for health telecare: Integration of LangChain, embeddings with OpenAI, and Pinecone using the question answering technique

O. Cárdenas[a]* • S. Falconí[b] • E. Tusa[a] • A. Rodríguez[a]

[a]*Universidad Técnica de Machala, Carrera de Tecnologías de la Información, Machala, Ecuador*
[b]*Universidad de Almería, Facultad de Ciencias Médicas, Almería, España.*

Received 11 10 2023; accepted 03 31 2024
Available 06 30 2024

Abstract: This study embarks on developing an innovative ChatBot model tailored for medical telecare, integrating state-of-the-art technologies like LangChain, OpenAI embeddings, and Pinecone to refine user-bot interactions through advanced question-answering techniques. Distinguished from direct ChatGPT usage, this ChatBot is uniquely designed to provide specialized medical advice, drawing upon a meticulously curated dataset from verified medical sources, ensuring unparalleled accuracy and reliability in responses. Utilizing Google Colab as the backbone for execution and intensive data processing, this research underscores the model's superior capacity in delivering context-specific, quality responses, thereby setting a new benchmark in telehealth communication. It delves into a comparative analysis of response times and cost-efficiency between GPT-3.5-turbo and GPT-4, revealing the strategic advantages of our ChatBot in enhancing healthcare delivery. Through this detailed exploration, the study showcases the pivotal role of customized ChatBots in elevating telecare services, marking a significant leap forward in the intersection of AI and healthcare.

Keywords: ChatBot, LangChain, Pinecone, OpenAI, health, care

*Corresponding author.
E-mail address: oecardenas@utmachala.edu.ec (O. Cárdenas).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

In the current context, digital health is in constant evolution, driven by technological advancements and the growing need for innovative solutions in healthcare. ChatBots, as emerging tools, promise to transform the interaction between healthcare professionals and patients (Sezgin et al., 2022). This research stems from the need to develop a robust and efficient ChatBot model that can be trained with verified data and utilize the most advanced tools available in the field of artificial intelligence. Various works have explored the application of ChatBots in healthcare, highlighting their potential to transform medical care and improve patient outcomes.

Various previous works have explored the application of ChatBots in the healthcare field. For instance, Sezgin et al. discuss the operationalization and implementation of artificial intelligence models, such as GPT-3, within the U.S. healthcare system, highlighting their potential to transform medical care. Another study conducted by Daniel et al. (2022) focused on the development of a ChatBot to answer hospital caregivers' questions about medications and pharmaceutical organization. Park et al. (2023) investigated the effect of a ChatBot's emotional disclosure on user satisfaction and the intention of reuse in the field of mental health counseling. Furthermore, Edeh et al. (2022) underscored the significance of machine learning algorithms in developing predictive models for diseases such as Hepatitis C.

Personalization in healthcare through ChatBots represents an area of innovation that can significantly improve user interaction and satisfaction. Recent research emphasizes how the personalization of responses and services by ChatBots can enhance the efficacy of healthcare, better adapting to individual patient needs (Sun & Zhou, 2023). On the other hand, as demonstrated by An et al., highlights the importance of using advanced technologies to improve care and analysis in the healthcare field (Tan et al., 2023).

The integration of technologies such as LangChain, embeddings, Pinecone, and GPT-3.5 represents a qualitative leap in the capability of these systems to process and generate precise responses. These tools, backed by solid theoretical-conceptual foundations, allow not only for a better understanding of natural language but also for the adaptability and continuous learning of the model. The present research is based on these concepts and seeks to advance the field by developing a ChatBot model that can be a benchmark in health Telecare.

2. Tools and techniques

2.1. Embeddings with OpenAI

Embeddings are vector representations that capture the semantics of entities such as words or phrases in a low-dimensional space. These representations have become essential in many natural language processing (NLP) tasks due to their ability to capture semantic and syntactic relationships between word (Ristoski et al., 2019). OpenAI, with its advanced models, has adopted and enhanced embedding techniques to provide more accurate solutions in NLP tasks (Ristoski & Paulheim, 2016).

RDF2Vec is an approach that employs language modeling techniques for the extraction of unsupervised features from word sequences and adapts them to RDF graphs. These sequences are generated by leveraging local information from graph substructures and learning latent numerical representations of entities in RDF graphs (Ristoski & Paulheim, 2016). Moreover, the integration of embedding techniques with neural translation has shown significant advances in reducing gender bias in machine translation (Escudé Font & Costa-Jussà, 2019). In addition, the combination of artificial intelligence with embedding tools and techniques has proven to be essential for ensuring privacy and security (Gupta et al., 2020).

The representation of words through vectors, also known as word embeddings (WE), has captured the attention of the natural language processing (NLP) field. These WE models can express syntactic and semantic similarities, as well as relationships and contexts of words within a given corpus. Although the most popular implementations of WE algorithms present low scalability, there are innovative approaches applying high-performance computing (HPC) techniques. Silva et al. proposed a wrapper library containing a set of optimizations to enhance the scalability and usability of these tools (da Silva et al., 2020).

2.2. LangChain

LangChain is presented as an innovative solution for the development of applications that benefit from advances in natural language processing (NLP). By providing modular components and standard chains, LangChain offers developers the flexibility to create customized applications that integrate with language models and other data sources (Ioannidis et al., 2023).

- **Data knowledge:** One of the most valuable features of LangChain is its ability to connect language models with other data sources. This feature is crucial as it allows applications to access and process information from various sources, improving the accuracy and relevance of the responses generated by the model.
- **Agentic interaction:** Allowing a language model to interact with its environment is crucial for dynamic applications. This interaction can include adapting to user inputs, integrating with other applications or systems, and the ability to perform actions based on user instructions.
- **Modular components:** Modularity is essential for the agile and flexible development of applications. By offering

components that are easy to use and customize, LangChain allows developers to adapt and expand their applications according to the specific needs of their project.

- Standard chains: These predefined chains provide a starting point for developers, facilitating the rapid implementation of common tasks. However, the true power of LangChain lies in its ability to customize and expand these chains according to the project's needs.

2.3. Pinecone

Pinecone is a vector database, which is a type of database designed to store and query high-dimensional vectors. Vectors are numerical representations of data, such as text, images, audio, and video.

Pinecone uses an indexing algorithm called vector quantization to store vectors in an index space. Vector quantization divides the index space into a set of cells, and each vector is assigned to the nearest cell (Query Data, n.d.).

This allows Pinecone to perform vector search queries efficiently. To search for a vector, Pinecone first finds the cell in which the vector resides. Then, it can query the cell to find other vectors that are nearby (Pinecone, n.d.).

2.4. Question answering technique

The QA technique has been developed to address challenges at the intersection of natural language understanding and information retrieval. Its aim is to interpret a question in natural language and search a database or dataset to provide an appropriate answer. Over the years, this technique has evolved and adapted to address diverse types of data, from text to images and videos.

- Visual reasoning: Datasets, like the one proposed by Hudson and Manning (2019), have been specifically designed for visual reasoning and compositional question answering. These datasets use scene graph structures to create questions that require diverse reasoning, and all questions come with functional programs that represent their semantics.
- Spatiotemporal reasoning: The extension of QA into the video domain has led to the introduction of tasks that require spatiotemporal reasoning. For example, the TGIF-QA dataset, introduced by Jang et al. (2017), is specifically designed for question answering in videos, where questions may require an understanding of temporal sequences and spatial relationships within a video.
- Integration of deep learning: With the advent of deep neural networks, QA systems have begun to integrate models such as convolutional neural networks (CNNs) to improve accuracy and reasoning capabilities. Noh et al. (2016) proposed a CNN-based approach with a dynamic parameter layer to improve accuracy in ImageQA tasks.
- Fundamental techniques: Various fundamental techniques underpin modern QA systems, especially when working with knowledge bases. (Diefenbach et al., 2018) highlight that these techniques combine methods from natural language processing, information retrieval, and semantic techniques to provide accurate answers to complex questions.

In conclusion, the question answering technique represents a field in constant evolution that combines advanced techniques from natural language processing, information retrieval, and machine learning to provide accurate responses to questions posed in natural language.

2.5. Objectives

Main objective:

Develop a ChatBot model for Telecare and health using innovative techniques and tools.

Specific objectives:

- Investigate and analyze current natural language processing and machine learning techniques that are relevant for the development of ChatBots in the health sector.
- Use LangChain for data processing and management, and Pinecone as a vector database to optimize search speed and accuracy.
- Integrate the question answering technique and embeddings with OpenAI to improve the accuracy and relevance of the responses provided by the ChatBot.
- Use the pre-trained language model GPT-3.5-turbo to process the responses and make them more user-friendly.

3. Development of the ChatBot model

To develop a ChatBot model, it is essential to have an appropriate development environment that facilitates the execution of the necessary libraries. In this context, Python has emerged as one of the most popular programming languages for natural language processing and ChatBot development. The use of Jupyter Notebook provides an interactive environment that allows writing and executing code, visualizing results, and documenting the development process (Baptista, 2021; Sitar & Leary, 2023).

Google Colab, on the other hand, is a platform that allows the execution of Jupyter Notebooks in Google's cloud without the need to set up a local environment. It is especially useful for those who want to access high-performance computational resources without incurring additional costs. Moreover, Google Colab facilitates real-time collaboration and access to powerful GPUs, which is beneficial for data-intensive processing tasks (Baptista, 2021).

A study highlights the use of Google Colab as an educational tool to teach physical chemistry using the Python programming language. This study emphasizes the ease of use of Google Colab, as it does not require the installation, configuration, and setting up of Python packages and libraries on personal computers (Baptista, 2021).

Another study highlights the integration of Jupyter with other tools and platforms, demonstrating the versatility and power of Jupyter Notebook for various applications (De Jesus Martinez et al., 2023).

3.1. Data collection for the dataset

Manual data collection from official medical sources is a meticulous and essential process to ensure the accuracy and reliability of the content that will feed the ChatBot. By obtaining information directly from authorized medical sources, it is ensured that the ChatBot provides evidence-based answers and current medical recommendations.

The manual collection process involves:

- **Identification of reliable sources:** This step involves selecting medical databases, specialized journals, recognized health organizations, and other sources known for their scientific rigor and accuracy. For this study, priority was given to official data from highly credible organizations such as the World Health Organization (WHO), the Pan American Health Organization (PAHO), and the Ministry of Health of Ecuador, thus guaranteeing access to firsthand verified procedures and information.
- **Information extraction:** Once the sources are identified, relevant information is extracted, including disease symptoms, recommended treatments, recovery times, precautions, among others.
- **Data organization:** The collected information is organized into specific categories or topics to facilitate its integration into the ChatBot model. For example, all information related to fractures could be organized under a "fractures" theme with subcategories like "symptoms", "treatments", "recovery time", etc.
- **Validation of information:** It is crucial to review and validate the collected information to ensure its accuracy. This may involve comparing the information against multiple sources or consulting with field experts.
- **Integration into the ChatBot model:** Once validated, the information is integrated into the ChatBot model. This may involve converting the information into a format that the ChatBot can understand and use to respond to questions.

It is important to note that manual data collection ensures that the ChatBot is fed with high-quality information, which is essential when dealing with health and well-being. By providing answers based on reliable medical sources, the ChatBot becomes a valuable tool for those seeking accurate and up-to-date medical information.

The use of ChatBots in the medical field has proven to be a valuable tool, especially when it comes to collecting patient data. For instance, a study highlights how a ChatBot helped patients with primary headache disorders through personalized text messages and how it was used to collect patient-reported outcomes (Chaix et al., 2022).

On the other hand, data collection in automated systems has proven to be a challenge due to the heterogeneity of systems, protocols, and interfaces. A study proposes a semi-automated system capable of extracting all kinds of data, including notes, to prepopulate clinical research forms (Trunzer et al., 2021).

Finally, in medical research, the traditional way of collecting data, that is, sifting through patient files, has been shown to induce biases, errors, and costs. A study presents a user-friendly solution to complete clinical research forms, reducing human effort and providing higher quality data (Quennelle et al., 2023).

3.2. Data handling and processing with LangChain using the chunking method

The "chunking" technique or chunk analysis in text processing refers to the division of text into smaller units (or "chunks") that have a coherent meaning. These chunks can be words, phrases, or even paragraphs. In the context of natural language processing (NLP), "chunking" is a form of shallow parsing that focuses on identifying patterns in the text without the need to analyze its complete grammatical structure (Wu & Liu, 2010).

In LangChain, chunking is the process of dividing a text document into smaller parts, called chunks. Chunks can vary in size and can be defined by different criteria, such as the number of words, the number of characters, or punctuation.

Chunking is an important part of natural language processing because it allows language models to work with more manageable text data. It can also help improve the performance of language models, as the models do not have to process the entire text document at once as represented in Figure 1.

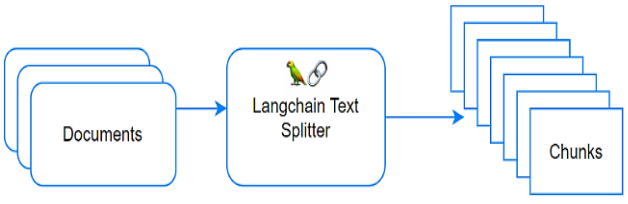

Figure 1. Document processing with text splitter.

In the study "Sentence Boundary Extraction from Scientific Literature of Electric Double Layer Capacitor Domain: Tools and Techniques", the importance of accurately extracting sentence boundaries from PDF documents for readability and natural language processing is highlighted. In this context, the term "chunk" refers to a block of text that needs to be processed and understood (Miah et al., 2022).

Figure 2 below demonstrates data processing with LangChain.

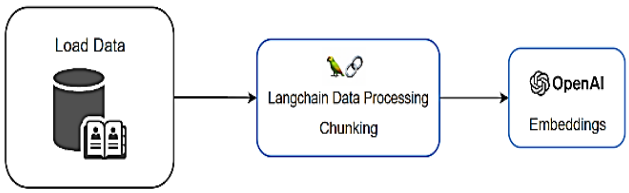

Figure 2. Data loading and chunk implementation for embeddings.

Load the data located in the Data folder and load them and by using the text splitter method, implement the chunks, where we can define the chunk size so that it can be processed and ready for the embedding method to be applied.

3.3. Processing with embeddings and storage in Pinecone

Once the data has been loaded and has gone through the process of LangChain and chunking, it is essential to transform them into a form that can be easily interpreted and processed by machine learning models.

Data vectorization, especially in the realm of natural language processing, is a crucial stage in transforming textual information into numerical representations that can be interpreted by machine learning models. embeddings, which convert words or phrases into high-dimensional vectors, have proven to be effective in capturing semantic and contextual relationships in the text (Kumari & Lobiyal, 2022). OpenAI's API offers advanced tools for generating high-quality embeddings, leveraging pretrained models and sophisticated deep learning techniques (Radford et al., 2019).

Once the data has been transformed into these vectors, it is essential to have an efficient system for storing and retrieving these embeddings. Pinecone, as a vector database, is designed to store and query high-dimensional vectors, providing fast and efficient retrieval, which is crucial for real-time applications such as a ChatBot (Johnson et al., 2021).

The integration of tools such as LangChain for initial processing, OpenAI for the creation of embeddings, and Pinecone for vector storage, establishes an optimized workflow for the development of a ChatBot focused on Telecare in health. This combination ensures that the ChatBot can provide evidence-based and up-to-date responses to user inquiries.

Figure 3 illustrates the previously explained concepts using the APIs of OpenAI and Pinecone.

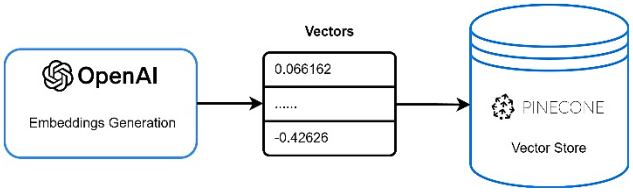

Figure 3. Vector storage process generated by embeddings.

The necessary libraries for these processes are imported, as well as the embedding of the APIs for their operation. In this process, the ChatBot is not yet executed; it is only the data processing.

In this process, the text that has already gone through the chunking process is called, sent to undergo the embedding process at OpenAI, and the result received is sent directly to the Pinecone vector database to have access to the vectors of the phrases at any time.

3.4. Sending of questions and obtaining results with the question answering method and application of GPT for response optimization

After the embedding process and vector storage, the model is ready to receive questions.

The question answering (QA) method is an advanced technique in natural language processing that focuses on providing accurate answers to specific questions based on a dataset or corpus (Rajpurkar et al., 2016). With the integration of GPT-3.5-turbo or GPT-4, a language model developed by OpenAI, the quality and relevance of the generated responses can be significantly improved (Brown et al., 2020).

Once the data has been processed and stored in Pinecone, the model is ready to receive and respond to questions. When a query is sent to Pinecone, it returns the most similar vectors based on cosine similarity or Euclidean distance. These vectors represent text snippets or information that are relevant to the question.

Figure 4 explains the process.

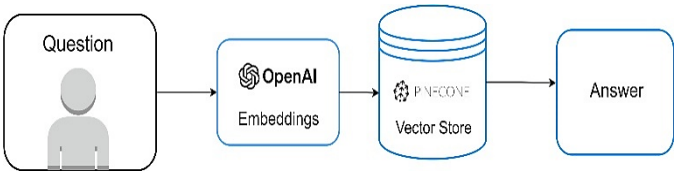

Figure 4. Questions and answers with Pinecone.

We enter the question, and then an embedding and vectorization process is conducted on the question to perform the calculation and mathematical comparison, returning the most matching results to the question. It will return the texts in raw form, sometimes with poor semantic structure.

However, as mentioned, the results returned by Pinecone can be raw texts with poor semantic structure. This is where GPT-3.5-turbo or GPT-4 comes into play. By feeding these text snippets into the model, the responses can be refined and optimized to be more coherent and relevant to the user.

Figure 5 demonstrates the response refinement process.

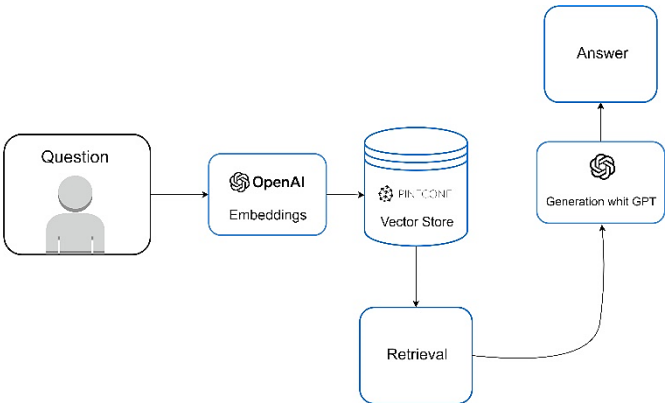

Figure 5. Question answering process applying LLM.

Next, a detailed description of the process is provided:

- The question is converted into vectors so that it can be sent for comparison to Pinecone and receive the most appropriate response.
- The raw response is returned and sent to the LLM for processing and grammatical and semantic improvement.
- The LLM returns the processed response to the ChatBot, which is then presented to the user.

This process ensures that the generated answers are relevant and precise, leveraging both the power of Pinecone for information retrieval and the capability of GPT-3.5-turbo or GPT-4 to generate coherent and well-formed responses.

4. Results

4.1. Data processing and storage

Data processing using the embedding technique was a crucial step in preparing medical information for use in the ChatBot. This technique transforms textual information into numerical representations, known as high-dimensional vectors. These vectors capture the essence and semantic context of the data, allowing the ChatBot model to interpret and respond to queries more accurately.

4.1.1. Vector generation with embeddings

- Each record of the medical dataset was processed to convert it into a vector.
- These vectors, despite being numerical representations, retain the crucial information and contextual relationships of the original text.
- Limitations: The cost of use can be a limitation since it is not a free tool, and the larger the amount of data, the higher the cost of use.

Knowledge of the tool is important for its best applicability; without a knowledge base, the difficulty increases.

- Dimensionality is the vector space of each chunk's embeddings; with this, we can see its value in space and represent it.

4.1.2. Storage in Pinecone

Once generated, the vectors were transferred to Pinecone, a specialized vector database.

Pinecone provides efficient mechanisms for storing and retrieving vectors, which is essential to ensure quick response times when the ChatBot is operational.

These vectors did not occupy significant storage space in Pinecone, but considering the dimension we configured, we have a maximum limit of vectors per metadata. In this case, we have a dimension of 1536, which allows us a maximum of 100,000 vectors.

In Figure 6, a 3D scatter plot and a 2D scatter plot can be observed to understand the vector storage space in Pinecone.

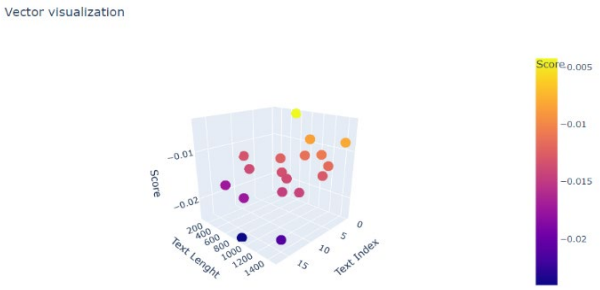

Figure 6. Graphic representation of vector database.

Clusters within the vector space are observed, these proximities reveal the closeness between phrases, which indicates the similarity between words and grammatical structure, forming clusters. We also see that the color ranges from warm to cool; this indicates that the more yellow the vector, the richer it is grammatically, hence it has a higher score, and the opposite occurs with lower vectors.

In Figure 7, we can better observe the classification of the score and groupings of words with similar vectors.

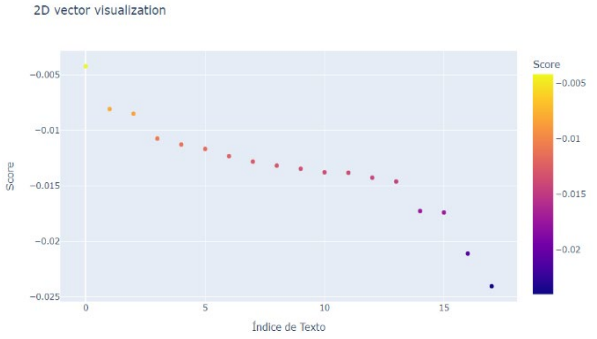

Figure 7. Score classification and vector grouping in Pinecone.

4.2. Representation of responses with the LLM

4.2.1. Response time analysis: Detailed evaluation of GPT-3.5-turbo and GPT-4

Evaluating the performance of language models in terms of response times is crucial, especially in critical contexts such as healthcare, where every second counts. A system that can provide rapid and accurate responses is vital for the efficient operation of the ChatBot and, therefore, for providing quality healthcare assistance.

Methodology

The evaluation of response times was conducted using a series of ten specific questions designed to reflect the diversity and complexity of medical inquiries that might arise in actual interaction with the ChatBot.

Question design: The questions were conceptualized to cover different medical areas and degrees of urgency. Questions regarding COVID-19, orthopedic surgeries, and postoperative care were included, thus reflecting the versatility required for an effective health ChatBot. The questions also varied in terms of specificity and context, ranging from symptoms and diagnoses to care recommendations and postoperative considerations, to evaluate the models' ability to respond appropriately and accurately to several types of queries.

Registration and analysis process: For each question, the response times from GPT-3.5 and GPT-4 models were precisely measured using the Gradio interface. This tool tracks the time from the moment a user submits a question to when an answer is received, providing a quantitative measure of response efficiency. This methodology emphasizes not only the rapidity of responses but also the quality and accuracy of the information provided. Special attention is given to the models' comprehension of the context and specificity of each question, assessing their ability to produce coherent and pertinent replies.

Example of questions:

- Area: COVID-19

"What are the most common symptoms of COVID-19?"

"If I have muscle pain and intense fatigue, could it be a symptom of COVID-19?"

- Area: Orthopedics – femur fracture

"After femur fracture surgery, how long does it typically take to recover?"

- Area: Postoperative care – ankle fracture

"After surgery for an ankle fracture, what precautions should I take when showering to protect the affected area?"

Supplementary questions used in the study:

This appendix provides a detailed overview of the specific questions used in the evaluation of the ChatBot models, GPT-3.5-turbo, and GPT-4. These questions were selected to cover a range of common medical inquiries, reflecting the diversity and complexity of questions a user might pose in actual healthcare scenarios. The questions aim to assess the models' ability to provide accurate, relevant, and clear information in response to common healthcare-related queries.

- Question 1: Symptoms of COVID-19

"If I have muscle pain and intense fatigue, could it be a symptom of COVID-19?" This question evaluates the models' ability to recognize and communicate potential symptoms of COVID-19, offering users valuable guidance on when to consider the possibility of an infection.

- Question 2: Severe symptoms of COVID-19

"What should I do if I experience severe symptoms of COVID-19, such as difficulty breathing and confusion?" This inquiry evaluates the models' capacity to provide critical advice on handling severe symptoms of COVID-19, emphasizing the urgency and appropriateness of the recommended actions.

- Question 3: Post-operative care for ankle fracture

"After ankle fracture surgery, what precautions should I take when showering to protect the affected area?" This question assesses the models' effectiveness in offering practical post-operative care advice, ensuring the safety and well-being of individuals recovering from an ankle fracture surgery.

- Question 4: COVID-19 vaccination

"What are the recommended COVID-19 vaccines for people with a history of severe allergic reactions?" This question aims to assess the models' knowledge on vaccine recommendations, particularly for individuals with specific health concerns.

- Question 5: Managing COVID-19 at home

"How should I care for myself at home if I have mild symptoms of COVID-19, such as fever and cough?" This inquiry evaluates the ability of the models to provide self-care advice for managing mild symptoms of COVID-19 at home.

- Question 6: Femur fracture recovery activities

"What types of physical activities are safe to engage in during the first two months of recovery from femur fracture surgery?" This question evaluates the models' guidance on safe physical activities post-femur fracture surgery to aid in recovery

luation, the response time for each question listwithout risking further injury.

- Question 7: Pain management after femur fracture surgery

"What are effective pain management strategies following femur fracture surgery?" Here, the focus is on evaluating how well the models can suggest pain management techniques for patients recovering from femur fracture surgery.

- Question 8: Monitoring for complications post-surgery

"What signs of complications should I watch for in the weeks following surgery for a femur fracture?" This question assesses the models' ability to inform patients about potential post-operative complications and when to seek medical attention.

- Question 9: Ankle fracture post-operative care

"How often should the dressing on an ankle fracture surgery wound be changed, and what are the signs of infection to look out for?" This inquiry evaluates the models' advice on wound care and infection prevention following ankle fracture surgery.

- Question 10: Physical therapy for ankle fracture

"When is it typically safe to start physical therapy after ankle fracture surgery, and what exercises might be included?" The question evaluates the models' recommendations on the timing and types of physical therapy exercises suitable for ankle fracture recovery.

Results

The visualization of data through graphs and tables provides an effective means for the visual and numerical interpretation of the response times of both models, GPT-3.5-turbo, and GPT-4, across a variety of medical questions. To ensure a precise evaed below represents the average time measured over 10 trials. This approach helps to account for variability and ensures the reliability of the results presented.

In Figure 8, a visual comparison between the response times of GPT-3.5-turbo and GPT-4 for each question is depicted. This chart intuitively highlights the differences between the two models, visually showing how GPT-4, in most cases, has longer response times than GPT-3.5-turbo.

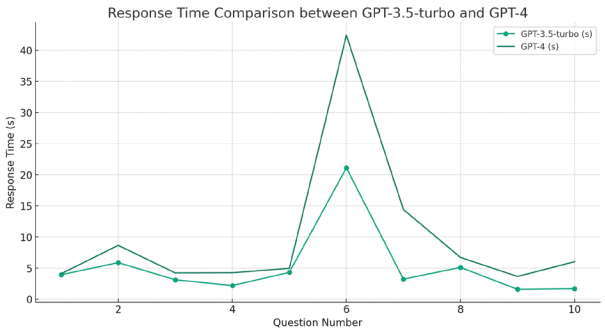

Figure 8. Demonstrative diagram of response time.

Figure 8 clearly illustrates the differences in response times between the two models for each question. It can be observed how the line representing GPT-4 is higher than the line representing GPT-3.5-turbo.

Variability and trends: Observing Table 1 and Figure 8, there is notable variability in the response times of both models. GPT-4 tends to have higher response times for most questions, with significant differences on questions like 6 and 7.

Consistency in quick responses: In certain instances, the difference in response times is minor, as seen with Question 1, suggesting that for certain types of questions, both models can offer responses in similar timescales.

Extreme times: Question 6 for GPT-4 reflects the longest response time, which might require further analysis to understand the reasons behind this anomaly. Conversely, GPT-3.5-turbo on Question 10 shows an extremely efficient response time.

Table 1. Response time of GPT language models.

Questions	GPT-3.5-turbo(s)	GPT-4 (s)
1	3.93	4.09
2	5.87	8.65
3	3.12	4.22
4	2.19	4.25
5	4.31	4.96
6	21.11	42.42
7	3.22	14.41
8	5.11	6.71
9	1.59	3.65
10	1.69	6.02

Importance of question complexity: The nature and complexity of the questions are crucial factors in interpreting these results. More elaborate questions could lead to longer response times.

Quality vs speed: Although this analysis focuses on response times, it is essential to also consider the quality of the responses provided by each model in future analyses, as a faster model is not necessarily more accurate or relevant.

4.2.2. Comparative cost analysis: Assessing the cost-effectiveness of GPT-3.5-turbo and GPT-4 in specific inquiries

The cost evaluation between GPT-3.5-turbo and GPT-4 is conducted to determine which model is more economically cost-effective in the context of specific medical inquiries and to understand how the cost structure of each model can influence the choice of implementation.

Cost calculation methodology

Costs have been calculated based on the number of tokens processed by each model to answer each of the specific questions. The cost per token for each model has been multiplied by the total number of tokens processed in each response.

Cost outcome

Here is Table 2, which shows the costs associated with each of the 10 questions for both models.

Table 2. Cost tabulation and tokens used per question.

Questions	GPT-3.5-turbo(usd)	Tokens	GPT-4(usd)	Tokens
1	$0.00125	828	$0.025	828
2	$0.00146	954	$0.030	949
3	$0.00183	1216	$0.037	1225
4	$0.00160	1057	$0.031	1045
5	$0.00180	1189	$0.036	1183
6	$0.00331	2025	$0.074	1979
7	$0.00188	1247	$0.047	1392
8	$0.00256	1678	$0.055	1705
9	$0.00245	1630	$0.049	1637
10	$0.00208	1384	$0.044	1420

Analysis of costs and tokens

The analysis of the cost and token usage for GPT-3.5-turbo and GPT-4 across ten questions reveals the following insights:

- Average cost:

GPT-3.5-turbo: $0.00203 USD

GPT-4: $0.04338 USD

- Average tokens:

GPT-3.5-turbo: 1320.8 tokens

GPT-4: 1336.3 tokens

- Total cost:

GPT-3.5-turbo: $0.02028 USD

GPT-4: $0.4338 USD

- Total tokens:

GPT-3.5-turbo: 13,208 tokens

GPT-4: 13,363 tokens

The visualization highlights a significant difference in cost per question between GPT-3.5-turbo and GPT-4 in Figure 9, with GPT-4 being more expensive. However, the token usage between the two models is similar, indicating that the higher cost associated with GPT-4 is not due to increased token consumption but reflects its enhanced capabilities and performance.

In summary, while GPT-4 offers potentially improved performance or features over GPT-3.5-turbo, this comes at a significantly higher cost, with only a slight increase in token usage. This information can be valuable for users considering the trade-off between cost and the benefits of using a more advanced model for their specific needs.

Token variation: The variation in the number of tokens processed for each question highlights a direct correlation with the costs in Figure 9. It is noteworthy that, even in cases where the number of tokens is similar, the costs for GPT-4 are significantly higher.

Practical implications: These cost differences can have substantial implications when choosing a model, especially for applications where operational costs are a crucial consideration.

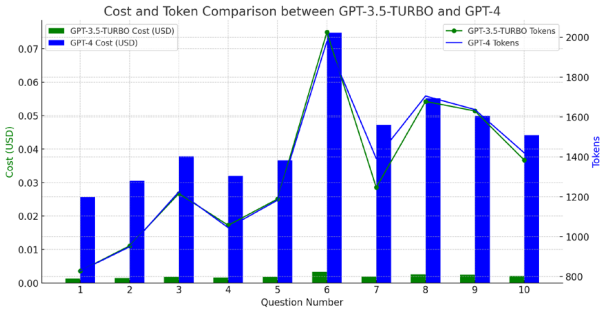

Figure 9. Difference in cost between natural language models.

The presented data show that GPT-4 is more expensive than GPT-3.5-turbo for all evaluated inquiries, despite similarities in the number of tokens processed for several questions. This cost difference suggests that, although GPT-4 may offer improvements in the quality and accuracy of responses, its implementation might not be cost-effective in all contexts, and should be carefully considered based on available resources and the specific needs of the application.

4.2.3. Assessment of response quality: Detailed analysis of GPT-3.5-turbo and GPT-4 on specific inquiries

In the evaluation of the quality of responses between GPT-3.5-turbo and GPT-4, precision, relevance, completeness, and clarity of the answers have been compared. These are crucial aspects in the medical field where the accuracy and detail of the information can significantly impact clinical decision-making and patient well-being.

Evaluation methodology

Each response from both models was evaluated on the following criteria, with scores assigned on a scale from 1 to 5:

- **Precision:** Matching with the information provided in the dataset.
- **Relevance:** The pertinence of the information provided in response to the question.
- **Completeness:** The extent to which the response addresses all aspects of the question.
- **Clarity:** How easily the response can be understood.

Results

The evaluation of the response quality of GPT-3.5-turbo and GPT-4 to common medical inquiries was conducted by healthcare professionals from the Technical University of Machala. This approach ensured rigor and reliability in the assessment, focusing on the models' performance according to four criteria: accuracy, relevance, completeness, and clarity. Each criterion was scored on a scale from 1 (poor) to 5 (excellent), reflecting the depth and utility of the models' answers to three specific questions detailed in Table 3, Table 4 and Table 5, thereby providing a structured framework for comparison.

Table 3. Evaluation of LLM with Question 1.

Question 1: If I have muscle pain and intense fatigue, could it be a symptom of COVID-19?		
GPT-3.5-turbo		
ACCURACY	4	It mentions that these are fewer common symptoms of COVID-19.
RELEVANCE	4	Relevant, but it could mention more common symptoms to provide context.
COMPLETENESS	4	Complete but could include more symptoms and recommendations.
CLARITY	5	Clear and concise.
GPT-4		
ACCURACY	4	It mentions that they are symptoms and recommends medical attention.
RELEVANCE	5	Relevant, especially due to the mention of having been in contact with someone with COVID-19.
COMPLETENESS	4	Informative, though it could be more comprehensive by mentioning more symptoms.
CLARITY	5	Clear and direct.

- The average rating for GPT-3.5-turbo is 4.25.
- The average score for GPT-4 is 4.5.

This indicates that, on average, GPT-4 has been evaluated slightly better than GPT-3.5-turbo in terms of accuracy, relevance, completeness, and clarity in the context of answering the question of whether muscle pain and severe fatigue can be symptoms of COVID-19.

- The average rating for GPT-3.5-turbo is 3.25.
- The average score for GPT-4 is 5.0.

This analysis reveals that GPT-4 has been evaluated significantly better than GPT-3.5-turbo in all aspects considered (accuracy, relevance, completeness, and clarity) in relation to the response on what to do when faced with severe COVID-19 symptoms, such as difficulty to breathe and confusion.

Table 4. Evaluation of LLM with Question 2.

Question 2: What should I do if I experience severe symptoms of COVID-19 such as difficulty breathing and confusion?		
GPT-3.5-turbo		
ACCURACY	3	Advises calling the healthcare provider but does not specify that these are severe symptoms that require immediate attention.
RELEVANCE	3	Relevant but does not specify the severity of the symptoms.
COMPLETENESS	3	Incomplete in terms of specifying the severity of the symptoms and the need for immediate medical attention.
CLARITY	4	Clear, but could be more specific.
GPT-4		
ACCURACY	5	It mentions the need for immediate medical attention and suggests contacting the healthcare provider.
RELEVANCE	5	Relevant and specific.
COMPLETENESS	5	Complete and specific regarding the need for immediate medical attention.
CLARITY	5	Clear and specific.

- The average rating for GPT-3.5-turbo is 3.0.
- The average score for GPT-4 is 4.5.

This analysis demonstrates that GPT-4 has been evaluated significantly better than GPT-3.5-turbo in terms of accuracy, relevance, completeness, and clarity regarding recommendations on precautions to take when showering after ankle fracture surgery.

According to the analysis results, GPT-4 tends to outperform GPT-3.5-turbo in terms of accuracy, relevance, completeness, and clarity. It provided more detailed and specific answers, which are vital in medical contexts.

In Figure 10, we can observe in detail how GPT-4 is superior to GPT-3.5-turbo.

Table 5. Evaluation of LLM with Question 3.

Question 3: After ankle fracture surgery, what precautions should I take when showering to protect the affected area?		
GPT-3.5-turbo		
ACCURACY	3	Provides a brief and general answer about protecting the area with a plastic bag.
RELEVANCE	3	Relevant but not detailed enough.
COMPLETENESS	2	Not sufficiently complete, lacking detail and other important considerations.
CLARITY	4	Clear but not complete.
GPT-4		
ACCURACY	4	Provides specific details on how to protect the affected area, mentions not to submerge the ankle in water, and gives recommendations to prevent accidents in the shower.
RELEVANCE	5	Completely relevant, detailed, and specific.
COMPLETENESS	4	Very complete, addresses various aspects of postoperative care during showering.
CLARITY	5	Clear and detailed.

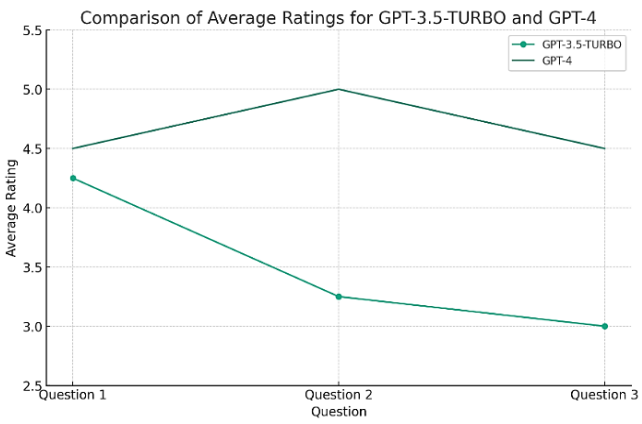

Figure 10. LLM response score.

5. Discussion of results

Data processing and storage

The use of the embeddings technique in our study has been fundamental for the preparation of medical information. This technique, which transforms textual information into high-dimensional vectors, has allowed for the capture of the essence and semantic context of the data, thereby improving the ChatBot's accuracy in interpreting and responding to inquiries. However, limitations such as the cost of usage and the need for deep knowledge of the tool for effective application were identified.

The storage of vectors in Pinecone has proven to be efficient, allowing for quick response times and optimizing storage space. The graphical representation of the vector database has revealed significant clusters in the vector space, indicating grammatical and structural similarities between the phrases.

Representation of responses with the LLM

The assessment of response times for GPT-3.5-turbo and GPT-4 has revealed significant differences between the two models. GPT-4, while superior in the quality and accuracy of responses, has shown higher response times, which could be a critical factor in healthcare contexts where every second counts. The variability in response times suggests the influence of the nature and complexity of the questions on the response times.

Comparative cost analysis

The cost analysis has revealed that GPT-4 is more expensive than GPT-3.5-turbo, despite similarities in the number of tokens processed. This difference in costs suggests that the choice of model should be carefully considered, evaluating profitability and available resources, especially in applications where operational costs are a crucial consideration.

Quality of responses assessment

In terms of response quality, GPT-4 has outperformed GPT-3.5-turbo in accuracy, relevance, completeness, and clarity, providing more detailed and specific answers, which are vital in medical contexts. However, GPT-4's superiority in quality must be weighed alongside response time factors and costs to determine its practical application viability.

Concluding thoughts and comparison with previous studies

The convergence of our results with previous research in the field of medical informatics reinforces the validity and relevance of our study. The implementation of embeddings in AI ChatBots and their effective visualization are crucial for the evolution of healthcare, allowing for more accurate and contextually rich diagnoses and assistance.

A study by Oniani and Wang (2020) explored the use of language models, specifically GPT-2, to automatically answer questions related to COVID-19, applying transfer learning and retraining the model on the COVID-19 open research dataset (CORD-19) corpus (Oniani & Wang, 2020). This study highlights

the importance of relevance and accuracy in responses generated by ChatBots in the health context, especially in pandemic situations where precise and up-to-date information is crucial. In their assessment, models like BERT and BioBERT outperformed others in relevance-based sentence filtering tasks, highlighting the importance of selecting appropriate models and techniques to enhance the response quality in medical ChatBots.

The comparison with previous studies, such as those by Beam et al. (2020) and Lee et al. (2021), validates the applicability and relevance of our ChatBot in the healthcare context, demonstrating the potential of AI systems to enhance patient-healthcare system interaction and provide accurate health assistance (Beam et al., 2020; Lee et al., 2021). The superiority of GPT-4 in response quality, as observed in our study, suggests a significant advancement in the language models' ability to generate more detailed, accurate, and contextually rich responses, albeit with important considerations in terms of costs and response times.

6. Conclusions

Data processing: The embeddings technique has been essential for transforming textual information into numerical representations, optimizing the ChatBot's interpretation and response.

Storage in Pinecone: It has allowed for efficient vector retrieval and has revealed significant clusters in the vector space, indicating grammatical and structural similarities between phrases.

Response time: GPT-4, though superior in quality and accuracy, has shown higher response times than GPT-3.5-turbo, which could be critical in urgent healthcare contexts.

Costs: GPT-4 is more expensive than GPT-3.5-turbo, necessitating careful consideration of cost-effectiveness and available resources when choosing the model.

Quality of responses: GPT-4 has outperformed GPT-3.5-turbo in terms of accuracy, relevance, completeness, and clarity, providing more detailed and specific responses in medical contexts.

7. Recommendations

Comprehensive evaluation: It is crucial to consider the factors of response quality, response times, and costs in an integrated manner when choosing the language model to implement in medical applications.

Resource optimization: A careful evaluation of available resources and the specific need of the application is recommended to determine the feasibility of implementing more advanced and costly models like GPT-4.

In-depth analysis: A deeper analysis of anomalies in response times and the influence of question complexity on response times and quality of responses is suggested.

Knowledge development: Developing a deep understanding of the tools used, such as the embeddings technique, is vital to optimize their applicability and overcome the identified limitations.

This analysis and discussion provide a solid foundation for future research and development in the implementation of language models in medical applications, considering critical factors such as response quality, response times, and costs.

Conflict of interest

The authors do not have any type of conflict of interest to declare.

Funding

The authors received no specific funding for this work.

References

Baptista, L. (2021). Using Python and Google Colab to Teach Physical Chemistry During Pandemic. [Preprint]. https://doi.org/10.26434/chemrxiv.13656665.v1

Beam, A. L., Kompa, B., Schmaltz, A., Fried, I., Weber, G., Palmer, N., ... & Kohane, I. S. (2020). Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. In *Pacific Symposium on Biocomputing* (Vol. 25, pp. 295-306). https://doi.org/10.1142/9789811215636_0027

Brown, T. B., 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:2005.14165). arXiv. http://arxiv.org/abs/2005.14165

Chaix, B., Bibault, J. E., Romain, R., Guillemassé, A., Neeral, M., Delamon, G., ... & Brouard, B. (2022). Assessing the performances of a chatbot to collect real-life data of patients suffering from primary headache disorders. *Digital Health*, *8*, 20552076221097783. https://doi.org/10.1177/20552076221097783

Daniel, T., de Chevigny, A., Champrigaud, A., Valette, J., Sitbon, M., Jardin, M., ... & Renet, S. (2022). Answering Hospital Caregivers' Questions at Any Time: Proof-of-Concept Study of an Artificial Intelligence–Based Chatbot in a French Hospital. *JMIR Human Factors*, *9*(4), e39102. https://doi.org/10.2196/39102

De Jesus Martinez, T., Hershberg, E. A., Guo, E., Stevens, G. J., Diesh, C., Xie, P., ... & Holmes, I. H. (2023). JBrowse jupyter: a Python interface to JBrowse 2. *Bioinformatics*, *39*(1), btad032. https://doi.org/10.1093/bioinformatics/btad032

Diefenbach, D., Lopez, V., Singh, K., & Maret, P. (2018). Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information systems*, *55*, 529-569. https://doi.org/10.1007/s10115-017-1100-y

Edeh, M. O., Dalal, S., Dhaou, I. B., Agubosim, C. C., Umoke, C. C., Richard-Nnabu, N. E., & Dahiya, N. (2022). Artificial intelligence-based ensemble learning model for prediction of hepatitis C disease. *Frontiers in Public Health*, *10*, 892371. https://doi.org/10.3389/fpubh.2022.892371

Escudé Font, J. E., & Costa-Jussà, M. R. (2019). Equalizing gender biases in neural machine translation with word embeddings techniques. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing, 147–154.* https://doi.org/10.18653/v1/W19-3821

Gupta, R., Tanwar, S., Al-Turjman, F., Italiya, P., Nauman, A., & Kim, S. W. (2020). Smart contract privacy protection using AI in cyber-physical systems: tools, techniques and challenges. *IEEE access*, *8*, 24746-24772. https://doi.org/10.1109/ACCESS.2020.2970576

Hudson, D. A., & Manning, C. D. (2019). GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6700-6709). https://doi.org/10.1109/CVPR.2019.00686

Ioannidis, J., Harper, J., Quah, M. S., & Hunter, D. (2023). Gracenote. ai: Legal Generative AI for Regulatory Compliance. In *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023)*. https://doi.org/10.2139/ssrn.4494272

Jang, Y., Song, Y., Yu, Y., Kim, Y., & Kim, G. (2017). Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2758-2766). https://doi.org/10.48550/arXiv.1704.04497

Johnson, J., Douze, M., & Jégou, H. (2021). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535-547. https://doi.org/10.1109/TBDATA.2019.2921572

Kumari, A., & Lobiyal, D. K. (2022). Efficient estimation of Hindi WSD with distributed word representation in vector space. *Journal of King Saud University-Computer and Information Sciences*, *34*(8), 6092-6103. https://doi.org/10.1016/j.jksuci.2021.03.008

Lee, H., Kang, J., & Yeo, J. (2021). Medical specialty recommendations by an artificial intelligence chatbot on a smartphone: development and deployment. *Journal of medical Internet research*, *23*(5). https://doi.org/10.2196/27460

Miah, M. S. U., Sulaiman, J., Sarwar, T. B., Naseer, A., Ashraf, F., Zamli, K. Z., & Jose, R. (2022). Sentence boundary extraction from scientific literature of electric double layer capacitor domain: tools and techniques. *Applied Sciences*, *12*(3), 1352. https://doi.org/10.3390/app12031352

Noh, H., Seo, P. H., & Han, B. (2016). Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR), (pp. 30-38). https://doi.org/10.1109/CVPR.2016.11

Oniani, D., & Wang, Y. (2020). A qualitative evaluation of language models on automatic question-answering for covid-19. In *Proceedings of the 11th ACM international conference on bioinformatics, computational biology and health informatics* (pp. 1-9). https://doi.org/10.1145/3388440.3412413

Park, G., Chung, J., & Lee, S. (2023). Effect of AI chatbot emotional disclosure on user satisfaction and reuse intention for mental health counseling: a serial mediation model. *Current Psychology*, *42*(32), 28663-28673. https://doi.org/10.1007/s12144-022-03932-z

Pinecone Docs. (n.d.). Pinecone Docs. Retrieved October 12, 2023, from https://docs.pinecone.io/home

Quennelle, S., Douillet, M., Friedlander, L., Boyer, O., Neuraz, A., Burgun, A., & Garcelon, N. (2023). The Smart Data Extractor, a clinician friendly solution to accelerate and improve the data collection during clinical trials. In *Caring is Sharing–Exploiting the Value in Data for Health and Innovation* (pp. 247-251). IOS Press.
https://doi.org/10.3233/SHTI230112

Query data. (n.d.). In Pinecone. Retrieved October 12, 2023, from https://docs.pinecone.io/docs/query-data

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.
https://api.semanticscholar.org/CorpusID:160025533

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392.
https://doi.org/10.18653/v1/D16-1264

Ristoski, P., & Paulheim, H. (2016). RDF2Vec: RDF Graph Embeddings for Data Mining. *International Workshop on the Semantic Web*.
https://api.semanticscholar.org/CorpusID:35288341

Ristoski, P., Rosati, J., Di Noia, T., De Leone, R., & Paulheim, H. (2019). RDF2Vec: RDF graph embeddings and their applications. *Semantic Web*, *10*(4), 721-752.
https://doi.org/10.3233/SW-180317

Sezgin, E., Sirrianni, J., & Linwood, S. L. (2022). Operationalizing and implementing pretrained, large artificial intelligence linguistic models in the US health care system: outlook of generative pretrained transformer 3 (GPT-3) as a service model. *JMIR medical informatics*, *10*(2), e32875.
https://doi.org/10.2196/32875

da Silva, M. L., Meyer, V., Kirchoff, D. F., Neto, F. J., Vieira, R., & De Rose, A. C. (2020). Evaluating the performance and improving the usability of parallel and distributed Word Embeddings tools. In *2020 28th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)* (pp. 201-206). IEEE.
https://doi.org/10.1109/PDP50117.2020.00038

Sitar, M. C., & Leary, R. J. (2023). Technical note: colab_zirc_dims: a Google Colab-compatible toolset for automated and semi-automated measurement of mineral grains in laser ablation–inductively coupled plasma–mass spectrometry images using deep learning models. *Geochronology*, *5*(1), 109-126.
https://doi.org/10.5194/gchron-5-109-2023

Sun, G., & Zhou, Y. H. (2023). AI in healthcare: navigating opportunities and challenges in digital communication. *Frontiers in Digital Health*, *5*, 1291132.
https://doi.org/10.3389/fdgth.2023.1291132

Tan, L., Tan, O. K., Sze, C. C., & Goh, W. W. B. (2023). Emotional Variance Analysis: A new sentiment analysis feature set for Artificial Intelligence and Machine Learning applications. *Plos one*, *18*(1), e0274299.
https://doi.org/10.1371/journal.pone.0274299

Trunzer, E., Vogel-Heuser, B., Chen, J. K., & Kohnle, M. (2021). Model-driven approach for realization of data collection architectures for cyber-physical systems of systems to lower manual implementation efforts. *Sensors*, *21*(3), 745.
https://doi.org/10.3390/s21030745

Wu, J., & Liu, L. (2010). Chunk Parsing and Entity Relation Extracting to Chinese Text by Using Conditional Random Fields Model. *Journal of Intelligent Learning Systems and Applications*, 02(03), 139–146.
https://doi.org/10.4236/jilsa.2010.23017