Innovation in Geriatric Care: An AI Assistant with LLM Integration Based on Health Guidelines

Juliana Basulo-Ribeiro[1][a,*], Nádia C. G. Matos[2,3][b], Sabrina Magalhães Araujo[2,3,4][c], Nuno Capela[2,3][d], Francisco Bischoff[2,4,5][e], Leonor Teixeira[1][f] and Ricardo Cruz-Correia[2,4,5][g]

[1]Department of Economics, Management, Industrial Engineering and Tourism (DEGEIT), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Intelligent Systems Associate Laboratory (LASI), University of Aveiro, Aveiro, Portugal

[2]Department of Community Medicine, Information and Health Decision Sciences (MEDCIDS), Faculty of Medicine, University of Porto, Porto, Portugal

[3]Health Data Science Ph.D. Program, Faculty of Medicine of the University of Porto, Porto, Portugal

[4]CINTESIS - Center for Health Technology and Services Research, Faculty of Medicine, University of Porto, Porto, Portugal

[5]VirtualCare, Portugal

Keywords: Large Language Model, Artificial Intelligence, Healthcare, Geriatric Care.

Abstract: With the advance of artificial intelligence and natural language processing technology, a new tool is standing out in the field of understanding and generating natural language in a sophisticated way: the Large Language Model (LLM). According to several authors, LLMs can be used for various types of medical cases, providing access to different sources of information, and have opened up countless opportunities in the healthcare sector. This work aims to share the lessons learned during the process of developing an LLM-based assistant aimed at specific pathologies that are more prevalent in the elderly in order to support caregivers, whether they are private individuals, home care organizations, nursing homes or others. This study has a significant potential impact on the community by providing access to detailed information on developing an LLM-based assistant.

1 INTRODUCTION

Language is vital for human communication and self-expression, as well as being fundamental for interaction between people and machines (Naveed et al., 2023). With the advance of artificial intelligence (AI) and natural language processing technology, a new tool is coming to the fore in the field of understanding and generating natural language in a sophisticated way: the Large Language Model (LLM). Language modeling is a long-standing research topic, dating back to the 1950s, and today, LLMs have evolved from statistical language models to neural language models, pre-trained language models and finally to large-scale LLMs (Minaee et al., 2024). Examples of LLMs include OpenAI's GPT (Generative Pre-trained Transformer), Google's T5 (Text-To-Text Transfer Transformer) and Google's BERT (Bidirectional Encoder Representations from Transformers) (Brown et al., 2020; Devlin et al., 2018; Raffel et al., 2019).

Smith et al. (2024) points out 10 rules for using LLMs in science, as these AI tools are set to change the way we operate in many areas: (1) Recognize that LLMs are powerful, but they can make mistakes; (2) Confirm the information generated by the models with

[a] https://orcid.org/0000-0002-3411-3519
[b] https://orcid.org/0009-0007-7841-6574
[c] https://orcid.org/0000-0003-1443-2106
[d] https://orcid.org/0009-0007-7841-6574
[e] https://orcid.org/0000-0002-5301-8672
[f] https://orcid.org/0000-0002-7791-1932
[g] https://orcid.org/0000-0002-3764-5158

Basulo-Ribeiro, J., Matos, N. C. G., Araujo, S. M., Capela, N., Bischoff, F., Teixeira, L. and Cruz-Correia, R.
Innovation in Geriatric Care: An AI Assistant with LLM Integration Based on Health Guidelines.
DOI: 10.5220/0013096100003911
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 18th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2025) - Volume 2: HEALTHINF*, pages 361-371
ISBN: 978-989-758-731-3; ISSN: 2184-4305
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

reliable sources; (3) Train the model with reliable data to ensure that it is representative and free of bias; (4) Clearly specify the context so that the answers are more accurate; (5) Be aware of the biases present in the data used to train the model, and take steps to mitigate them; (6) Collaborate and share knowledge, experiences and challenges to achieve best practices for preparing LLMs; (7) Maintain transparency by disclosing the methodology, tools and limitations when presenting work that incorporates LLMs; (8) Consider the ethical implications of using language models; (9) Use models as a support tool, but not as a substitute for critical thinking and rigorous scientific methods; (10) Actively collaborate to improve language models and promote responsible and ethical AI practices.

Today, we can already imagine a future world where health care is monitored not only by physicians and nurses but also by solutions that integrate AI, capable of understanding, interpreting, and even predicting health needs. This is the world where LLMs can play an extremely vital role in healthcare. Behind the scenes of the digital revolution, these models are shaping a new era of healthcare, where communication between patients and healthcare professionals reaches unprecedented levels of personalization and efficiency. This is a topic discussed in the literature, where Mehandru et al. (2024) mention in their study that: "Recent developments in large language models (LLMs) have unlocked opportunities for healthcare, from information synthesis to clinical decision support.". LLM can help to improve patient care and medical diagnosis (Nassiri & Akhloufi, 2024)

According to several authors, LLMs can be used for various types of medical cases, giving access to different sources of information and tools, including clinical guidelines, databases with electronic health records, among others, which shows the possibility and importance of developing LLMs for health based on guidelines (Mehandru et al., 2024; Park et al., 2023; Yang et al., 2023). Smith et al. (2024) point out that LLMs have the potential to transform science, but they also present significant challenges and risks. Thus, evaluating LLMs becomes crucial in all contexts; however, it is in the medical context that evaluation is most critical (Nassiri & Akhloufi, 2024). The accuracy, reliability and effectiveness of LLMs are essential to ensure that the information provided can be used safely and effectively in healthcare, and in an environment where decisions can have a significant impact on people's lives, rigorous evaluation of LLMs becomes not only necessary but indispensable (Mehandru et al., 2024; Yang et al., 2023). Nassiri and Akhloufi (2024) mention the importance of training the model using specific health documents, all to reduce the risk of producing incorrect information. A significant ethical concern with the use of these models is the risk of perpetuating biases and inaccuracies in medical data and information, which can severely impact patient care (Nassiri and Akhloufi, 2024). As Nassiri and Akhloufi (2024) point out, there are many "technical and ethical issues that need to be resolved before LLMs can be used extensively in the medical field".

It is crucial to emphasize the need for a collaborative and multidisciplinary approach to this type of AI-based technology, with interaction between different specialists, such as: technology professionals, healthcare professionals, and health regulators. This collaboration not only promotes innovation in healthcare but also ensures that these technologies are implemented responsibly, maximizing patient benefits and minimizing potential risks. (Cascella et al., 2024; Piñeiro-Martín et al., 2023; Thapa & Adhikari, 2023)

Population aging is one of the most significant demographic phenomena of the 21st century, and as the elderly population grows, so do the challenges associated with managing health conditions specific to this age group. Among these conditions, non-oncological pathologies are particularly prevalent and require special attention from both caregivers and healthcare professionals. Often, a lack of knowledge about these diseases and the best practices for preventing or managing them results in complications and a deterioration in the quality of life (Maresova et al., 2019; Sun & Li, 2023; World Health Organization, 2015, 2022).

This paper aims to present a project to develop an assistant based on LLMs, which uses health guidelines as a basis for supporting the care of the elderly. This initiative follows the recommendations proposed by Nassiri and Akhloufi (2024) in their study, mentioned earlier in this section. Consequently, this article is structured as follows: section 2 describes the practical case carried out and the discussion; and at the end, section 3 presents the final remarks, which include the conclusion, the contributions of this work, the limitations and future work for the advancement of LLMs.

2 PRACTICAL STUDY

2.1 Goals and Methods

This article analyses the knowledge acquired during the development of an assistant designed to deal with specific non-oncological pathologies that commonly affect the elderly. The main objective is to provide adequate support to carers, whether they are private individuals, home care organisations, nursing homes or others. In addition, this paper explores the challenges encountered in developing a prototype of a large language model (LLM) adapted to the care of the elderly. It also seeks to understand how the development process can reveal wider obstacles that any research can face when working with scientific documentation, including the specificities of LLMs and the training techniques required for their effective application.

To analyse the lessons learnt during the development of the assistant, a qualitative method was adopted, based on a documentary review and critical reflection on the team's experiences throughout the project.

The methodology was divided into three main stages:

i. Data collection: Review of documents and records related to the LLM development project, including planning documents and reports/documents of findings over time. In addition, informal interviews were conducted with team members involved in developing the assistant to gain insights into the challenges faced, strategies adopted, and lessons learned.

ii. Data Analysis: Organization and categorization of the data collected, highlighting the main areas of learning identified during the development process.

iii. Summary and Discussion: A summary of the analysis results, highlighting the main lessons learned throughout the development process. In addition, a critical discussion was held on the implications of these learnings for future projects of this scope and practice in general.

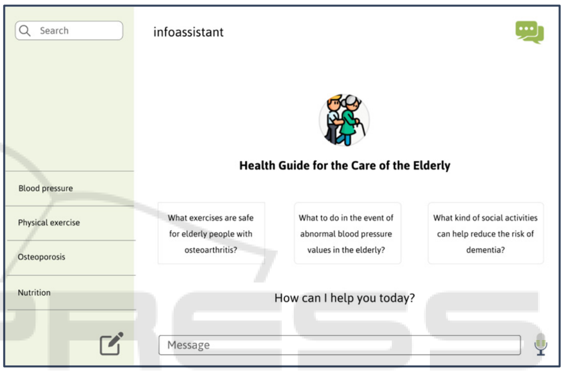
Health professionals and technology developers were involved in this process to promote this project's effective and responsible development.

2.2 Results

2.2.1 Description of the Assistant

The work developed, an assistant based on an LLM, is an innovative tool designed to support caregivers of the elderly by providing detailed information on prevalent diseases and medical guidelines. It uses different datasets from the Transparency Portal and INE Pordata, which can be updated every six months and every year. The model acts as a conversational assistant, able to answer questions and offer specific educational material aimed at the most common pathologies and care among the elderly. Among the main beneficiaries are nursing homes, parish councils, pharmacies, patient associations and home care providers.

To guarantee the accuracy and relevance of the information, this research was based on reliable sources such as the World Health Organization (WHO), PubMed and Elsevier, which ensures the project's high technical feasibility. The tool's innovation is evident, not only in the way it transforms complex data into useful answers but also in its ability to offer caregivers continuous support 24/7.



Health Guide for the Care of the Elderly

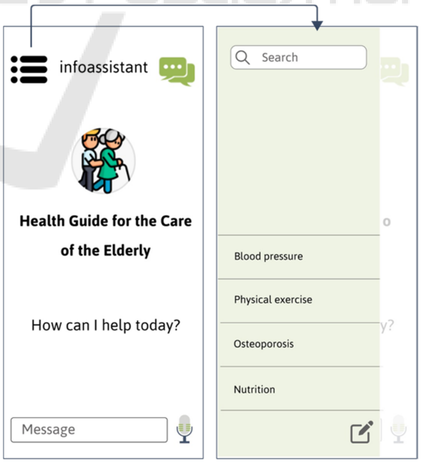
What exercises are safe for elderly people with osteoarthritis?

What to do in the event of abnormal blood pressure values in the elderly?

What kind of social activities can help reduce the risk of dementia?

How can I help you today?

Figure 1: LLM "infoassistant" desktop mock-up.



Health Guide for the Care of the Elderly

How can I help today?

Figure 2: LLM "infoassistant" mobile mock-up.

The tool's relevance to external entities is also highlighted, as it facilitates the dissemination of knowledge and good practices, contributing to improving care provided to the elderly. This combination of innovation, technical feasibility and positive impact makes this project a valuable solution in the field of geriatric care. Figure 1 and Figure 2 show a prototype of the natural language conversation assistant that we intend to make available at the end of the project in desktop and mobile versions, respectively.

2.2.2 Approach to Preparing the Assistant

This subsection aims to present the action plan conducted from the decision to develop this project to its preparation and availability. The plan for this project was divided into several work packages (WP) to make it easier to manage and execute them, allocating the different resources to each one.

WP1: Data Collection (Search and Download of Guidelines)→ The first WP in the initial phase, focused on searching for and downloading health guidelines related to non-oncological pathologies prevalent in the elderly population, using specialized websites such as PubMed. These documents, which are generally in PDF format, constitute the database for the project. At a later stage, this task was automated using Python technology, allowing relevant documents to be identified, downloaded and stored periodically. This systematic process, in line with the requirements identified, reduces the likelihood of errors and does not require human intervention. The methodology involved creating specific queries, processing the results, extracting relevant data and storing the information. Figure 3 illustrates the workflow of this WP, with the 'Search' stage representing the process of searching for guidelines using MeSH (Medical Subject Headings) terms; the 'Access to Research Databases' stage returns the results of the previous task and identifies the most relevant documents. The public API (Application Programming Interface) of the NCBI Entrez system (https://www.ncbi.nlm.nih.gov/ books/NBK25501/) was utilized, providing access to a wide range of Entrez databases, including PubMed, PMC, Gene, Nuccore and Protein. The third stage, 'Download and Store', represents the process of downloading and storing the guidelines, as well as extracting their metadata. Finally, the 'Structuring Metadata' stage aims to organize the metadata according to the structure defined for the inputs of the next work package, WP2.

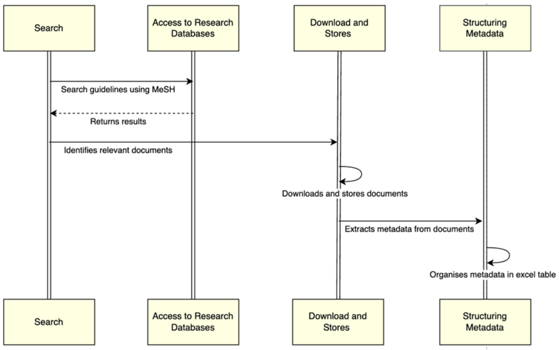

Search
Access to Research Databases
Download and Stores
Structuring Metadata
Search guidelines using MeSH
Returns results
Identifies relevant documents
Downloads and stores documents
Extracts metadata from documents
Organises metadata in excel table
Search
Access to Research Databases
Download and Stores
Structuring Metadata

Figure 3: Schematic representation of WP1.

WP2: Processing the Guidelines→ The second WP aims to analyse and convert the PDF documents collected (during the WP1 process) into text (questions and answers) to extract specific information from these documents. Group the guidelines by disease, considering the International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) system and for each disease coding, symptoms, signs, abnormal findings, complaints, social circumstances and external causes of injury or disease. With this set of information, a range of questions and answers was identified, which would serve as the basis for the following WP.

WP3: Develop and Refine the AI Assistant→ Considering that the assistant was designed to answer questions about non-oncological pathologies prevalent in the elderly population, based on the guidelines collected and processed in the previous work packages, this stage involved setting up a workspace in the AnythingLLM (AnythingLLM, 2024) tool, a framework to facilitate designed to facilitate interaction with natural language models, delivering precise answers. Various approaches were explored and combined during this stage, such as fine-tuning, transfer learning, RAG (Retriever-Augmented Generation), or even a combination of these techniques. These methods, defined in detail later in this document, allowed for the refinement and adjustment of the assistant to ensure it could provide useful and accurate answers to questions related to non-oncological pathologies in the elderly population based on the guidelines collected and processed. The outcome of this WP is an assistant capable of delivering relevant and precise responses. Another output of this WP is an API that allows other systems to access the developed assistant.

WP4: Making the Assistant Available→ The last WP of the project involved developing an interface that makes the assistant available, allowing users to interact directly with it. This interface would invoke the API developed in the previous WP to allow the assistant to be incorporated into another system, making it possible to clarify doubts, always based on the health guidelines collected, thus providing a useful tool for both caregivers, professionals in the field and even the elderly themselves.

Table 1 summarizes this information, and Figure 4 illustrates it, considering the sequence of events over the 4 WPs.

Table 1: Summary of each work package.

Work Packages	Objective	Considerations/ Outputs
WP0: Project Management		
WP1: Search and download guidelines	i.find/identify the guidelines; ii.download the guidelines; iii.Put the guidelines into a paste; iv.Generate a list of documents in the folder with information about the guidelines; v.Extra task: Identify the main illnesses and needs of the elderly within the project's scope.	• Folder with guidelines; • List with Timestamp, DOI, Name, URL • Various formats: .pdf, .doc, .txt
WP2: Edit the guidelines	i.Define the variables to extract; ii.Extract the data (the result is a table); iii.Translate the data into a question/answer format (table).	• List of data by guideline - 1st phase in EN • Context: Population; gender, ethnicity, age, etc... • Decision algorithm: what the guideline proposes as a solution • Question/Answer format - 1st phase in EN
WP3: Develop and refine the AI assistant	Use the WP2 data and refine LLM with the questions/answers. The output/result is a service that, when invoked, establishes the dialog.	• API to send a question and receive a reply • API for dialog

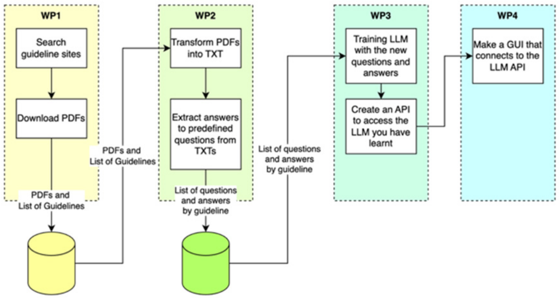

WP1
Search guideline sites
Download PDFs
PDFs and List of Guidelines
PDFs and List of Guidelines
WP2
Transform PDFs into TXT
Extract answers to predefined questions from TXTs
List of questions and answers by guideline
List of questions and answers by guideline
WP3
Training LLM with the new questions and answers
Create an API to access the LLM you have learnt
WP4
Make a GUI that connects to the LLM API

Figure 4: Schematic illustration of the LLM development project.

2.2.3 Lessons Learned

This section aims to share lessons learned throughout the development process of this project. By 'lessons learned,' we mean the challenges identified during the execution of each WP of the project and the methodologies that were applied to address them, including both the strategies that proved effective in solving the problem and those that did not achieve the expected results.

WP1: Search and Download Guidelines

- ***Efficient Research***

Creating specific queries using MeSH terms ensures that all relevant documents are identified, but failure to improve them can jeopardize the quality and relevance of the results returned.

- ***Data Processing***

The conversion of content in .pdf format to text (i.e., encoding, special characters), tables, flowcharts and figures sometimes leads to a lack of precision in the information extracted, which can be accentuated by different document formats and structures. This lack of precision, although not completely unknown, takes on greater relevance when we talk about the health sector since inaccurate information can lead to advice and/or decision-making that jeopardizes the health and safety of patients. To mitigate this situation, it is important to develop efficient algorithms to extract and organize the necessary metadata automatically, minimizing human intervention and errors. In addition, it is necessary to define an effective method for segmenting the data, such as the Chunking method (Dozza et al., 2013). The method must divide the content in such a way as to maintain the connection between data that shares the same context; for example, when extracting information from a health document, the algorithm must automatically identify and group together sections relating to the same patient, diagnosis or treatment, ensuring that the information extracted is accurate and contextualized.

- ***Storage and Organization***

Storing the guidelines as well as their metadata in an organized and accessible way proved to be a major challenge, due to the periodic updating of the data without redundancies, maintaining the integrity and timeliness of the information.

- ***Tokens***

Another challenge faced was the fact that many research databases use tokens in the process of downloading guidelines, and access tokens usually have a limited lifetime, expiring after a certain period. Automating this flow becomes complicated as this requires tokens to be obtained and renewed frequently. In addition, there may also be usage limitations, i.e., limitations on the number of requests that can be made within a certain period, which requires careful management to ensure that the defined limits are not exceeded. Thus, automating tasks that depend on tokens adds a layer of complexity to the data collection process. Managing the validity and security of tokens, automating renewal and preparing the system for potential changes to third-party systems are essential strategies for mitigating these types of challenges.

WP2: Edit the Guidelines

- ***Additional Section in the Guidelines***

Since health guidelines do not have a standard structure, we propose adding a section with specific fields to improve their usefulness and make them easier for algorithms to understand. By incorporating these additional fields, the models could more easily interpret the health guidelines, allowing them to be better trained. This would ensure access to reliable information in a clear and effective way. These fields could include:

o Summary: A summary of the main recommendations and conclusions of the guideline, allowing a quick understanding of its content.

o Content Overview: A more detailed description of the topics covered in the guideline, highlighting the areas of focus and the main information provided.

o Purpose and Usefulness: A clear explanation of the objectives of the guideline and how it can be used in clinical practice or even in the reformulation of health policies.

o Who is it for: An identification of the health professionals, managers, researchers or other stakeholders for whom the guideline has been developed, helping to ensure that it is aimed at the right audience.

o Structure of the Guideline: A description of how the guideline is organized, including main sections, subsections and how to navigate through the document.

o Frequently Asked Questions: A list of common questions about the guideline and its answers, addressing concerns that users may have when implementing its recommendations.

- ***Versions in the Guidelines***

In addition, these guidelines should have clearly identified versions so that algorithms and models can easily check whether they are using the latest version in order to ensure maximum effectiveness and relevance of the content provided. Including a versioning system would help ensure that the information and recommendations are always up to date, reflecting the most up-to-date information. This approach would also make it possible to verify changes over time, making it easier to understand evolutions and updates in the guidelines. To do this, the guidelines could have a section or be accompanied by a document where it is possible to have a summary of what has been added/changed/removed, a "release note". This would not only make it easier to compare the different versions but would also help answer questions about changes in procedures.

WP3: Train the AI Assistant

- ***Versions of LLM Models and Software***

Version management is crucial in software and LLM models to maintain a clear timeline of its development and the functionalities that make it up. Each version should record changes in functionality, performance and bugs fixed. This makes it possible to compare different versions to assess performance improvements directly. This practice helps maintain and update systems, provides transparency for end users, and facilitates the reproduction of results in scientific research.

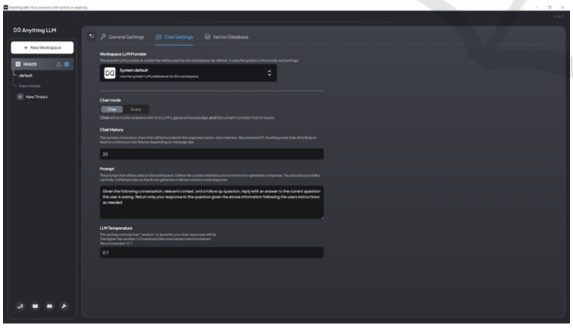
- ***Tool Used to Develop the Wizard***

For the project's development, the AnythingLLM (AnythingLLM, 2024) tool was chosen because it is an open-source platform offering significant flexibility for customization and integration into several projects. This solution supports various language models, including proprietary models, such as GPT-4, and open-source models, such as Llama and Mistral (https://useanything.com/). It includes a graphical user interface that facilitates interaction with the language model, integration with various data sources used to train and refine the models, and customization based on the specific needs of your application, all easily and intuitively. With AnythingLLM, there is no need to worry about privacy issues, a relevant point given that we are in the health sector, as it already implements measures to ensure that user data is treated securely and privately.

- ***Settings at AnythingLLM***

Within the AnythingLLM tool, there are several optional parameters that, while not mandatory, should be adjusted or defined, such as the model to be used (explained in the next few topics), to customize the model's behavior based on the specific needs of each project. Below are some of the settings available in the desktop version that can impact the model's behavior. Figure 3 shows a screenshot of the workspace settings.

- o <u>Chat Mode:</u> The "Chat" option, when active, sends the model the entire context of the conversation, i.e., the history of the conversation in the window. The "Query" option sends only the question asked by the user. Although context is important, it also contributes to performance degradation, so evaluating the objective and finding a balance is important.
- o <u>Chat History:</u> This is the number of questions/answers that will be sent as a history to the model whenever we are in "Chat" mode.
- o <u>Prompt: </u>Here, we should put relevant information and the instructions, in order to contribute to improving the performance of our assistant. For example, give a summary of the content of the documents that will serve as a knowledge base for the model we are using.
- o <u>LLM Temperature:</u> This parameter relates to the creativity of the LLM. It is set to a value between 0 and 1. The higher the value, the more creative the model's responses will be. By default, the software uses a value of 0.7, which is recommended by most models. In the test phase, we used lower values since we are in the health field. However, the answers provided by the model were very literal, lacking the personalization and naturalness of natural language. For this reason, we opted to restore the default value.



Workspace LLM Provider

The specific LLM provider & model that will be used for this workspace. By default, it uses the system LLM provider and settings.

System default

Use the system LLM preference for this workspace.

Chat mode

Chat Query

Chat will provide answers with the LLM's general knowledge and document context that is found.

Chat History

The number of previous chats that will be included in the response's short-term memory. Recommend 20. Anything more than 45 is likely to lead to continuous chat failures depending on message size.

20

Prompt

The prompt that will be used on this workspace. Define the context and instructions for the AI to generate a response. You should to provide a carefully crafted prompt so the AI can generate a relevant and accurate response.

Given the following conversation, relevant context, and a follow up question, reply with an answer to the current question the user is asking. Return only your response to the question given the above information following the users instructions as needed.

LLM Temperature

This setting controls how "random" or dynamic your chat responses will be.
The higher the number (1.0 maximum) the more random and incoherent.
Recommended: 0.7

0.7

Figure 5: Workspace settings for the desktop version of AnythingLLM.

- ***Testing the Different Existing Models***

The ability to quickly test different LLM models is essential for iterating and improving the quality of proposed solutions. Tools such as LM Studio (Element Labs, 2023) make it possible to conduct these tests quickly and intuitively, offering user-friendly interfaces and detailed analysis of model performance locally (on the machine). These tools can be configured to run a series of tests, such as assessing language comprehension, generating text and the ability to answer questions, which provides important metrics on the usefulness and quality of each model tested. All of this makes it possible to make better choices about which model(s) to use, depending on the purpose you want to achieve since the models are built for specific purposes, which makes them more accurate for the subjects they were developed for.

Another issue with models is that you must bear in mind that there are models that only perform well and efficiently in a specific language (e.g., EN) because they have been built and trained in that language and are, therefore, not multilingual. That is why it is important to test different models so that you can choose based on your use case.

- ***Choosing a Model***

At an early stage of the project, a document was drawn up with the pros and cons of the different models available, such as Falcon - Technology Innovation Institute (TII) in Abu Dhabi; Ihama 2 - Meta; Bloom; MPT-7B - MosaicML; Vicunha-13B - LMSYS ORG; GPT-3 - OpenAI; BERT; among others. However, after choosing the tool, an exhaustive analysis was also carried out of all the existing models in the Desktop version of AnythingLLM up to the date of development (May 2024), such as: LLaMA3 (Uncensored); LLaMA2 13B; LLaMA 3; CodeLlama 7B; Mistral 7B; Mistral 8x7B; Gemma 2B; Gemma 7B; Phi-2; Orca-Mini; Orca-Mini 7B; Orca-Mini 13B; among others.

When choosing the model to use in LLM to generate answers based on health guidelines regarding the care of the elderly and the most common pathologies in that age group, several factors need to be taken into account, such as: the size of the model, the ability to handle domain-specific information and whether or not the model has been optimized for health applications, as they may not have the controls or training necessary to handle sensitive and accurate medical/health information. In addition, it is important to ensure that any model used complies with local and international health data privacy regulations, such as GDPR and HIPAA, where applicable. Given the information found, it was decided to choose the **Orca-Mini 13B** model.

- ***Orchestrating the Solution***

Orchestration in LLM solutions involves managing and automating various tasks and services that comprise the LLM infrastructure, such as model training, testing, and model deployment. Using orchestration software like Flowise (Flowise, 2024) allows you to manage these tasks efficiently. In this case, Flowise was used for this purpose. This type of solution also allows different models to be used in the same solution, with questions being directed to the model with the best performance for the specific task. This ensures that the solution takes full advantage of the different capabilities of each model, maximizing the accuracy and efficiency of the answers provided.

- ***Model Training***

An unclear concept in this type of project is the issue of model training. In the context of this project, the term "training" will be used in a broad sense, referring both to the adjustment of model parameters (as in Fine-Tuning) and to the adaptation of the model through other techniques and approaches, which will be detailed later in this document. Since a pre-trained model was used, it was necessary to understand which techniques and/or approaches were used to adapt the model to the dataset, the guidelines.

Fine-Tuning is a technique that can be applied in these situations where pre-trained models are used. It involves further training the model using a specific dataset to adjust and optimize its performance for a particular task. This is achieved by adjusting the model's parameters with an additional specific dataset, the guidelines. Fine-tuning the base model with specific data from the collected health guidelines ensures that the model understands and responds to the nuances and specificities of these guidelines. To implement this technique, machine learning frameworks such as PyTorch or TensorFlow are recommended for adjusting the model. Libraries such as Hugging Face Transformers are also very helpful in this process.

Another technique is Transfer Learning, where the model pre-trained on a specific task is adapted to another related task. Instead of training the model from scratch, the knowledge gained during the initial training is reused, saving time and computational resources.

Retrieval-Augmented Generation (RAG), although not a training technique in the traditional sense, is a powerful approach to improving the quality of answers generated by language models during inference, especially in question-and-answer systems. It combines information retrieval and text generation. Rather than relying solely on the language model to generate answers, this approach first retrieves relevant documents from a database and then uses this information to generate more accurate and contextually relevant answers. This enhances the accuracy and relevance of responses, as the model does not just rely on its prior training but also uses up-to-date, context-specific information, ensuring that answers are always based on the most relevant documents by retrieving these guidelines before generating an answer. When combined with training techniques such as Fine-Tuning and Transfer Learning, RAG can contribute to a highly effective and accurate assistant, capable of providing informed and relevant answers based on specific guidelines.

It is therefore important to understand that “training” an LLM in this project involves several key considerations, and the choice of the appropriate techniques and approaches depends on various factors, including the project's objectives, available resources and the type of data. Below is a table summarizing the techniques and their applicability.

Table 2: Summary of training techniques and their applicability.

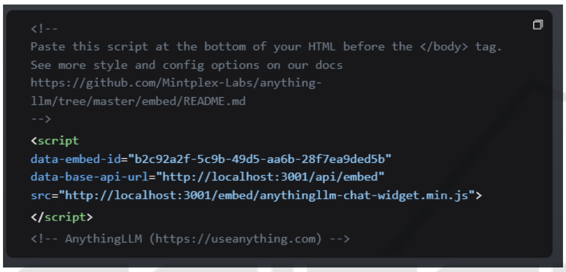
Technique	Description	Application
Transfer Learning	Use a pre-trained model as a starting point and adjust it for the specific task.	• When you have a limited set of specific data. • To save time and computing resources by using pre-trained models.
Fine-Tuning	Further train a pre-trained model with a specific data set to improve performance on a specific task.	• When you have domain-specific data and want to adapt a pre-trained model to improve accuracy and relevance.
Supervised Learning	Training a model with labeled data, with the goal of learning a function for mapping inputs to outputs.	• When you have a large set of labeled data. • For tasks such as text classification, answering specific questions, or sentiment analysis.
Unsupervised Learning	Training a model with unlabeled data to find hidden patterns or structures.	• When the data is not labeled. • For tasks such as clustering or dimensionality reduction.
Semi-Supervised Learning	Combining labeled and unlabeled data to improve model performance.	• When you have a limited amount of labeled data but a large amount of unlabeled data. • To increase model accuracy with less labeling effort.
Reinforcement Learning	Training a model where they learn to make decisions through rewards and punishments.	• For tasks that involve a sequence of decisions.

WP4: Make Assistant Available

- ***Summoning the Assistant***

Considering the specific characteristics of the AnythingLLM framework, it is important to note that invoking the wizard is only possible when using the contained version of the tool. Generally, the desktop version is used by a single user or in an individual development environment, where there is no room for the scalability and management of multiple instances required in a productive environment. So, given that the aim was to enable multi-user system execution and to "publish" the workspace on websites or other platforms, we decided to replicate the development process described in WP3 with the Docker version of the tool. Throughout the process, it was possible to see some differences, both in terms of the settings available and in terms of customization, with greater flexibility without significant impacts in terms of complexity for the user.

Another point is that the tool incorporates a chat widget that allows the workspace and its built-in knowledge base to be readily displayed on a website.



```
<!--
Paste this script at the bottom of your HTML before the </body> tag.
See more style and config options on our docs
https://github.com/Mintplex-Labs/anything-
llm/tree/master/embed/README.md
-->
<script
data-embed-id="b2c92a2f-5c9b-49d5-aa6b-28f7ea9ded5b"
data-base-api-url="http://localhost:3001/api/embed"
src="http://localhost:3001/embed/anythingllm-chat-widget.min.js">
</script>
<!-- AnythingLLM (https://useanything.com) -->
```

Figure 6: An example of a script tag embeds.

2.3 Discussions

This article aims to report on work in progress to create a health assistant based on guidelines that support an LLM. This effort corroborates the idea put forward by several authors that it is possible to create virtual health assistants based on reliable information (Mehandru et al., 2024; Park et al., 2023; Yang et al., 2023).

Advances in innovative LLM technology have shown significant potential to transform medical practice by providing quality and accessible care to healthcare professionals and patients, as noted by several authors in the literature (Mehandru et al., 2024; Nassiri & Akhloufi, 2024). However, due to the sensitive and critical nature of the healthcare sector in which they operate, it is imperative that these assistants are rigorously assessed at a clinical level. In their study, Nassiri and Akhloufi (2024) highlight the importance of this clinical assessment for any AI-based healthcare tool. In line with this thinking, our project recognizes the need to carry out this assessment to validate the development of this healthcare assistant.

Another relevant point to highlight is the need to continually train the model with new data and updated guidelines to maintain the assistant's relevance and accuracy, a critical component for the long-term success of any AI tool in healthcare.

Finally, it is worth noting the importance of multidisciplinary collaboration between technology developers, healthcare professionals and regulatory authorities to ensure that the development and implementation of LLM-based healthcare assistants are carried out in a responsible manner, as mentioned by several authors in the literature (Cascella et al., 2024; Piñeiro-Martín et al., 2023; Thapa & Adhikari, 2023). This collaborative approach safeguards patient safety and well-being.

3 FINAL REMARKS

3.1 Conclusion

Today, significant progress has been made in the field of LLMs. This progress results from the growing demand for innovative solutions to combat many daily challenges. These advances aim not only to improve the performance of LLMs in challenging tasks but also to explore new possibilities for previously unexplored applications, which will have a transformative impact on various areas of knowledge.

The final product resulting from this work, which will be presented to users, is more like a movie than a static image due to the dynamic nature of LLMs. Unlike an image, which remains the same over time, LLMs are designed to learn and adapt through use and interaction with users. This allows for continuous improvement of the responses it provides.

In developing the assistant presented in this study, the processes were initially conducted manually, reflecting a more exploratory initial phase. As the project progressed, there was a natural evolution towards automation. This "maturation" was evidenced by the progressive implementation of automated systems in each WPs. This transition made it possible to: optimize efficiency, reduce the time and effort required, and increase the accuracy and consistency of the results achieved.

3.2 Contributions and Implications

This study has a significant potential impact on the community by providing access to detailed information on how to prepare an assistant based on LLMs. This resource is extremely useful today, not only in the health area but in all sectors of activity. This LLM is an important tool to help people in their daily lives, facilitating access to relevant health information quickly and effectively.

3.3 Limitations and Future Work

One of the limitations of this study is that it has not been clinically validated. Therefore, in future work, we would like to evaluate LLM clinically. The current stage of development serves as an opportunity to conduct tests and consequent improvements, which will guarantee the effectiveness of the assistant when it is introduced to the market.

ACKNOWLEDGEMENTS

This work was supported by Portuguese funds through the Institute of Electronics and Informatics Engineering of Aveiro (IEETA) (UIDB/00127/2020) funded by national funds through FCT - Foundation for Science and Technology. This work has also been done under the scope of – and funded by – the Health Data Science Ph.D. Program of the Faculty of Medicine of the University of Porto, Portugal – heads.med.up.pt. The authors are thankful to the Program and its faculty for investing in their students and funding the open-access publication of their research. Additionally, this article was developed under the Mobilizing Agenda No. 41 - HfPT – HEALTH FROM PORTUGAL, co-funded by the PRR and the European Union through the Next Generation EU mechanism.

REFERENCES

AnythingLLM. (2024). *AnythingLLM*.

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*. http://arxiv.org/abs/2005.14165

Cascella, M., Semeraro, F., Montomoli, J., Bellini, V., Piazza, O., & Bignami, E. (2024). The Breakthrough of Large Language Models Release for Medical Applications: 1-Year Timeline and Perspectives. In *Journal of Medical Systems* (Vol. 48, Issue 1). Springer. https://doi.org/10.1007/s10916-024-02045-3

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. http://arxiv.org/abs/1810.04805

Dozza, M., Bärgman, J., & Lee, J. D. (2013). Chunking: A procedure to improve naturalistic data analysis. *Accident Analysis and Prevention*, *58*, 309–317. https://doi.org/10.1016/j.aap.2012.03.020

Element Labs, Inc. (2023). *LM Studio* (4).

Flowise. (2024). *Flowise*.

Maresova, P., Javanmardi, E., Barakovic, S., Barakovic Husic, J., Tomsone, S., Krejcar, O., & Kuca, K. (2019). Consequences of chronic diseases and other limitations associated with old age - A scoping review. *BMC Public Health*, *19*(1). https://doi.org/10.1186/s12889-019-7762-5

Mehandru, N., Miao, B. Y., Almaraz, E. R., Sushil, M., Butte, A. J., & Alaa, A. (2024). Evaluating large language models as agents in the clinic. In *npj Digital Medicine* (Vol. 7, Issue 1). Nature Research. https://doi.org/10.1038/s41746-024-01083-y

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). *Large Language Models: A Survey*. http://arxiv.org/abs/2402.06196

Nassiri, K., & Akhloufi, M. A. (2024). Recent Advances in Large Language Models for Healthcare. *BioMedInformatics*, *4*(2), 1097–1143. https://doi.org/10.3390/biomedinformatics4020062

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A Comprehensive Overview of Large Language Models*. http://arxiv.org/abs/2307.06435

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023, October 29). Generative Agents: Interactive Simulacra of Human Behavior. *UIST 2023 - Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. https://doi.org/10.1145/3586183.3606763

Piñeiro-Martín, A., García-Mateo, C., Docío-Fernández, L., & López-Pérez, M. del C. (2023). Ethical Challenges in the Development of Virtual Assistants Powered by Large Language Models †. *Electronics (Switzerland)*, *12*(14). https://doi.org/10.3390/electronics12143170

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. http://arxiv.org/abs/1910.10683

Smith, G. R., Bello, C., Bialic-Murphy, L., Clark, E., Delavaux, C. S., Lauriere, C. F. de, Hoogen, J. van den, Lauber, T., Ma, H., Maynard, D. S., Mirman, M., Lidong, M., Rebindaine, D., Reek, J. E., Werden, L. K., Wu, Z., Yang, G., Zhao, Q., Zohner, C. M., & Crowther, T. W. (2024). Ten simple rules for using large language models in science, version 1.0. *PLoS Computational Biology*, *20*(1). https://doi.org/10.1371/journal.pcbi.1011767

Sun, X., & Li, X. (2023). Editorial: Aging and chronic disease: public health challenge and education reform. *Front Public Health*, *16*(2), 107–118. https://doi.org/10.3389/fpubh.2023.1175898

Thapa, S., & Adhikari, S. (2023). ChatGPT, Bard, and Large Language Models for Biomedical Research: Opportunities and Pitfalls. In *Annals of Biomedical Engineering* (Vol. 51, Issue 12, pp. 2647–2651). Springer. https://doi.org/10.1007/s10439-023-03284-0

World Health Organization. (2015). *World report on ageing and health*.

World Health Organization. (2022). *Ageing and health*.

Yang, H., Yue, S., & He, Y. (2023). *Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions*. http://arxiv.org/abs/2306.02224.