Behaviour & Information Technology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tbit20

Incorporating evidence into mental health Q&A: a novel method to use generative language models for validated clinical content extraction

Ksenia Kharitonova, David Pérez-Fernández, Javier Gutiérrez-Hernando, Asier Gutiérrez-Fandiño, Zoraida Callejas & David Griol

To cite this article: Ksenia Kharitonova, David Pérez-Fernández, Javier Gutiérrez-Hernando, Asier Gutiérrez-Fandiño, Zoraida Callejas & David Griol (03 Mar 2024): Incorporating evidence into mental health Q&A: a novel method to use generative language models for validated clinical content extraction, Behaviour & Information Technology, DOI: 10.1080/0144929X.2024.2321959

To link to this article: https://doi.org/10.1080/0144929X.2024.2321959

Published online: 03 Mar 2024.

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2024.2321959

Incorporating evidence into mental health Q&A: a novel method to use generative language models for validated clinical content extraction

Ksenia Kharitonova[a], David Pérez-Fernández[b], Javier Gutiérrez-Hernando[a], Asier Gutiérrez-Fandiño[c], Zoraida Callejas[a,d] and David Griol[a]

[a]Department Software Engineering, University of Granada, Granada, Spain; [b]Department of Mathematics, Universidad Autónoma de Madrid, Ciudad Universitaria de Cantoblanco, Madrid, Spain; [c]LHF Labs, Bilbao, Bizkaia, Spain; [d]Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, Granada, Spain

ABSTRACT

Generative language models have changed the way we interact with computers using natural language. With the release of increasingly advanced GPT models, systems are able to correctly respond to questions in various domains. However, they still have important limitations, such as hallucinations, lack of substance in answers, inability to justify responses, or showing high confidence with fabricated content. In digital mental health, every decision must be traceable and based on scientific evidence and these shortcomings are hindering the integration of LLMs into clinical practice. In this paper, we provide a novel automated method to develop evidence-based question answering systems. Powerful state-of-the-art generalist language models are used and forced to employ only contents in validated clinical guidelines, tracking the source of the evidence for each generated response. This way, the system is able to protect users from hallucinatory responses. As a proof of concept, we present the results obtained building question-answering systems circumscribed to the clinical practice guidelines of the Spanish National Health System about the management of depression and attention deficit hyperactivity disorder. The coherence, veracity, and evidence supporting the responses have been evaluated by human experts obtaining high reliability, clarity, completeness, and traceability of evidence results.

ARTICLE HISTORY
Received 9 September 2023
Accepted 10 February 2024

KEYWORDS
Question-answering; generative language model; large language model; retrieval-augmented generation; clinical practice guide; mental health

1. Introduction

In recent years, the Internet and digital applications have led to an unprecedented surge in unstructured human language data. This is particularly true in the mental health domain with books of medical knowledge, scientific publications, patient histories, self reports, transcriptions of therapeutic conversations, etc.

This influx necessitated the development of novel approaches for rapid and efficient processing. Neural language models (Brown et al. 2020; Devlin et al. 2019; Radford et al. 2019; Touvron, Lavril, et al. 2023; Touvron, Martin, et al. 2023) built upon the Transformer's architecture (Vaswani et al. 2017), have emerged as the most effective solution for tackling a diverse array of natural language processing (NLP) tasks, such as sentiment analysis, text classification, machine translation, summarisation, conversational interaction, question-answering, and text generation.

Language models process word sequences (or tokens) to predict what comes next. Large language models (LLMs), with their vast size and extensive training data, demonstrate emergent abilities that surpass those of typical artificial intelligence models (Bowman 2023; Kaplan et al. 2020) and their use have been explored recently for a myriad of applications in the healthcare domains, such as medical education (Arif, Munaf, and Ul-Haque 2023; Eysenbach 2023; Sallam 2023), diagnosis (Arif, Munaf, and Ul-Haque 2023), treatment and patient care (Garg et al. 2023), clinical decision support (Liu et al. 2023a, 2023b), clinical research and development (Dahmen et al. 2023), compiling clinical data (DiGiorgio and Ehrenfeld 2023), carrying out standard tasks, such as creating of discharge summaries (Patel and Lam 2023) and clinical vignettes (Benoit 2023), as well as translating medical instructions into easier to understand language (Scerri and Morin 2023), among others.

CONTACT Zoraida Callejas zoraida@ugr.es Department Software Engineering, University of Granada, Avda. del Hospicio, Granada 18071, Spain; Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, Granada, Spain

Table 1. Comparison of our proposal with the main state-of-the art approaches.

Approach	General NLU capabilities	Domain-specific NLU capabilities	Hallucinations	Provides evidence for responses	Need to train / finetune
Use a general purpose pretrained model	High	Restricted	Potentially	No	No
Use a domain-specific pretrained models	Restricted	High (may be moderate for the specific downstream task)	Potentially	No	No
Fine-tune a general purpose model to the specific domain	High	High	Potentially	No	Yes
Our approach: Use a general purpose pre-trained model and force it to employ a trustworthy knowledge base in the specific domain	High	Restricted	No	Yes	No

LLMs may be developed in several ways. One option are pre-trained general-purpose models, e.g. OpenAI's GPT models, designed to comprehend and generate text across a wide spectrum of subjects and contexts. These models have a broad understanding of language but might lack specialised expertise in specific domains such as mental health. On the other end of the spectrum, LLMs can be trained within a particular domain, equipped with specialised knowledge and vocabulary. For instance, a language model trained solely on mental health literature would be able to understand and generate mental health texts with precision, but may have less sophisticated natural language reasoning capabilities and struggle to understand language outside their specialised domain. Another option is to fine-tune general-purpose LLMs with domain-specific data, allowing them to acquire nuanced understanding within that domain while retaining its broader linguistic capabilities. However, they may also suffer from overfitting when they are excessively tailored to the training data or be affected by catastrophic forgetting, when the tuning provokes that the model forgets some of its original knowledge and capabilities (Chronopoulou, Baziotis, and Potamianos 2019).

In this paper, we present a framework for developing question answering systems based on off-the-shelf pre-trained large language models. Pre-trained models have been exposed to diverse language patterns and contexts, enabling them to understand questions and generate responses very proficiently. Also they make system development more rapid, as they are readily available for use, and require less computational resources.

However, they may pose potential risks and have important limitations, particularly for the medical and mental health domain. As described in Harrer (2023), these models have been trained over a vast amount of usually unfiltered data, and consequently they may have ingested poor quality information, misinformation and falsehoods. Thus, when generating answers, they may reproduce or even amplify misleading, false or inappropriate contents and there is no way for the model to evaluate the generated answer or warn the user about it.

To harness the language comprehension and generation capabilities of generalistic pre-trained LLMs while mitigating associated risks, we propose an innovative approach, based on Retrieval-Augmented Generation (Gao et al. 2024), that selectively utilises their linguistic aptitude, isolating it from their topic knowledge. Instead, their outputs are produced exclusively on the basis of highly reliable knowledge bases. Moreover, we present a mechanism to ensure that these models refrain from generating responses unless they can confidently produce them. Additionally, the systems developed following our approach consistently provide verifiable textual excerpts, serving as sources of evidence for their generated responses. This meticulous evidence empowers users to trace and validate the origins of system outputs, thereby enhancing the system's transparency and explicability.

Table 1 presents a comparison of our proposal with other approaches. It is important to emphasise that our question-answering approach is not directly comparable to general-purpose LLMs: we exploit their sophisticated natural language understanding capabilities, but make them reason over our own trusted knowledge base instead of taking their responses. Thus, the knowledge from which it sources the answers is completely external to the LLM, which selects potential paragraph candidates for the answer and reasons through these paragraphs to generate the final response.

The rest of the paper is organised as follows. In Section 2 we present the state of the art and contextualise our proposal in the current advances in the field of language modelling and its application in mental health. Then, Section 3 describes our proposal, which have been used to build several Q&A systems as explained in Section 4. These systems have been evaluated, Section 5 presents a description of the evaluation method and discusses the results. Finally, Section 6 identifies the limitations of our proposal, Section 7 presents the main security and privacy aspects, and Section 8 presents the conclusions reached and points out several lines for future work.

2. State of the art

AI and machine learning have been used in the medical field for several decades. Surveys at the beginning of the 2000s, such as the one provided by Kononenko (2001), highlight the use of early statistical techniques such Bayesian networks, decision trees, and basic neural networks. These methods have evolved into more advanced methodologies based on natural language processing using deep learning techniques and LLMs (Ge et al. 2023; Yousaf et al. 2023). Within this field, language processing has been used for a very broad set of applications ranging from entity recognition, Liu, Zhou, and Wang (2019), to the development of neural network-based architectures to emulate the diagnostic process based on patients' symptoms (Richens, Lee, and Johri 2020).

LLMs have the potential to offer several benefits when applied to the field of medicine. These models can assist in analysing and extracting insights from vast amounts of medical data, including research papers, patient records, and clinical trial data. They can identify patterns, relationships, and relevant information that might be challenging for humans to uncover efficiently. LLMs can also quickly scan through and summarise a large volume of medical literature, helping healthcare professionals stay up-to-date with the latest research and evidence-based practices. This can save time and enhance the quality of decision-making.

One of the examples of this application of LLMs is their ability to synthesise knowledge by automatically generating secondary content, such as literature reviews (Taylor et al. 2022). These systematic reviews constitute an essential tool of Evidence Based Medicine (Haidich 2010) and are considered the strongest form of evidence to inform medical practice (Murad et al. 2016). Previous work has used LLMs to streamline the production of systematic reviews of the medical literature, for example, by helping to identify relevant studies (Lee et al. 2020; Miwa et al. 2014) or by automating data extraction (Gu et al. 2021; Wallace et al. 2016).

These models can also be used to assist in the diagnostic process by analysing patient symptoms, medical history, and test results. LLMs can aid researchers in generating hypotheses for further investigation by analysing existing data and suggesting potential research directions. They can suggest potential diagnoses and provide explanations for these suggestions based on existing medical knowledge. LLMs can offer evidence-based recommendations to healthcare providers by analysing patient data, medical guidelines, and treatment options. This can help clinicians make more informed decisions. In the case of mental health, they have been applied to predict mental health state from the analysis of online text data, with specific fine-tuned models being developed such as MentalBERT (Ji et al. 2021) or Mental-Alpaca.[1]

LLMs can generate patient-friendly explanations for mental health conditions, treatments, and procedures. This can improve patient understanding and adherence to recommendations, leading to better patient outcomes. It is also possible to create interactive and personalised educational materials for medical students, healthcare professionals, and patients. They can simulate medical scenarios, answer questions, and provide detailed explanations of complex concepts. LLMs can provide information about medications, including dosages, side effects, and potential interactions with other drugs. This can help prevent medication errors and enhance patient safety.

LLMs can also power natural language interfaces for medical applications, making it easier for users to interact with electronic health records (EHRs), medical databases, and health-related applications using everyday language and conversational AI systems. These models can also support telemedicine by offering virtual consultations and providing preliminary advice to patients, especially in scenarios where immediate access to a healthcare professional is limited. In addition, LLMs can assist in developing advanced natural language processing (NLP) tools specific to medical applications, such as sentiment analysis of patient feedback, automated transcription of medical conversations, and more. LLMs can also contribute to public health campaigns by generating clear and accessible information about disease outbreaks, preventive measures, and vaccination campaigns. This aids in disseminating accurate information to the general population.

The usefulness of LLMs in the field of mental health has gained increasing recognition in recent years (Cabrera et al. 2023; Callejas and Griol 2021; Lamichhane 2023). These models can generate easily understandable and relatable information about various mental health conditions, symptoms, treatment options, and coping strategies. More than 30 per cent of the world's population suffers from one or more mental health disorders; about 75 per cent of individuals in low- and middle-income countries and about 50 per cent of individuals in high-income countries do not receive care and treatment (Arias, Saxena, and Verguet 2016). The sensitive and often stigmatised nature of mental health conditions further exacerbates this problem, as many people find it difficult to openly disclose that they suffer from them (Corrigan and Matthews 2016; McCullock and Scrivano 2023).

LLMs can contribute to reducing the stigma associated with seeking mental health support. By providing a discreet and private channel for discussing mental health concerns, individuals may feel more comfortable addressing their issues. These models can also provide empathetic responses and emotional support to individuals in distress. While not a substitute for human interaction, LLMs can offer immediate comfort and validation, especially in moments when a person needs someone to talk to. In addition, they can engage users in conversations that encourage self-reflection and insight into their emotions, behaviours, and thought patterns. This process can aid individuals in gaining a better understanding of their mental health and identifying areas for personal growth.

In the particular case of question answering systems, LLMs are making a rapid progress and have renewed interest in the possibilities of AI for medical question answering tasks, a long-standing key challenge in this field (Nori et al. 2023; Singhal, Azizi, et al. 2023). Most of these approaches involve smaller linguistic models trained with domain-specific data, such as BioLinkBert, DRAGON, PubMedGPT, PubMedBERT, Med-PaLM 2 or BioGPT. However, the benchmarks achieved by these models have been substantially improved in a span of a few months using larger general-purpose LLMs such as GPT-3 and Flan-PaLM.

Levine et al. (2023) evaluated the diagnostic and triage accuracy of GPT-3 for 48 validated cases of common and serious conditions and compared them with those of lay-people and physicians. Additional studies have extended this work to evaluate GPT models performance in genetics, surgery, ophthalmology, etc. Antaki et al. (2023) and Duong and Solomon (2023). More recently, Ayers et al. (2023) have compared ChatGPT and physician responses on 195 randomly drawn patient questions from a social media forum and found ChatGPT responses to be rated higher in both quality and empathy.

However, as warned in Singhal, Tu, et al. (2023), before being used in real-world applications there is a need to warrant safety and validate their responses. Our strategic approach aims to seamlessly integrate LLMs into clinical practice, capitalising on the potential of the technology while mitigating potential risks. To progress in this direction, we identify significant advantages in concentrating on application domains where expert clinicians and mental health professionals are the main users, and technology is a complement which responses can be contextualised in the light of the professionals' experience. This is only possible if the answer is interpretable.

As explained in Yang et al. (2023) state-of-the-art responses generated by LLMs are seldom accompanied by justifications or supporting information sources. This undermines their usability as they can fabricate facts and produce false responses in a very confident manner. To tackle this issue, the concept of Retrieval-Augmented Generation (RAG) was introduced. This method limits the model's knowledge sources to verified databases supplied by the creators of the systems. At its core, RAG introduces an initial information retrieval phase, during which LLMs consult an external database to collect relevant information prior to beginning their question-answering or text generation tasks. This method not only shapes the subsequent text creation stage but also ensures that the generated responses are supported by the collected evidence, markedly increasing the accuracy and relevance of the final content (Gao et al. 2024).

Our proposal based on RAG serves as a step towards addressing the issues of hallucinated and false responses, as it allows developing explainable question answering systems based on trusted medical guidelines content.

3. Our proposal

The main idea behind our proposal is to leverage the sophisticated capabilities of current LLMs in question answering systems coupling them with an evidence-backed knowledge base. This way, LLMs are only used for language reasoning omitting the response that would be generated using their own knowledge base, which could be prone to problems such as hallucinations, lack of veracity and reduced interpretability. Instead, we restrict the contents considered to a reliable source. In our case, the clinical practice guidelines of the Spanish National Health System related to mental health (see Section 4.1).

幻覺

Figure 1 shows the main elements in our proposal.

Firstly (step 1), the knowledge base (KB) is created from the clinical guides (see Section 4.1) from their paragraphs and their corresponding embedding representations, which are computed using OpenAI API's improved second generation model `text-embedding-ada-002`.[2]

User queries are converted to embeddings using the same KB embedding representation (step 2). Subsequently, as step 3, we perform a semantic index search of the query embedding within the knowledge embedding database. Candidate answers, always extracted from clinical practice guidelines, are related in the embedding space to the question formulated by the user. This procedure generates a collection of candidate

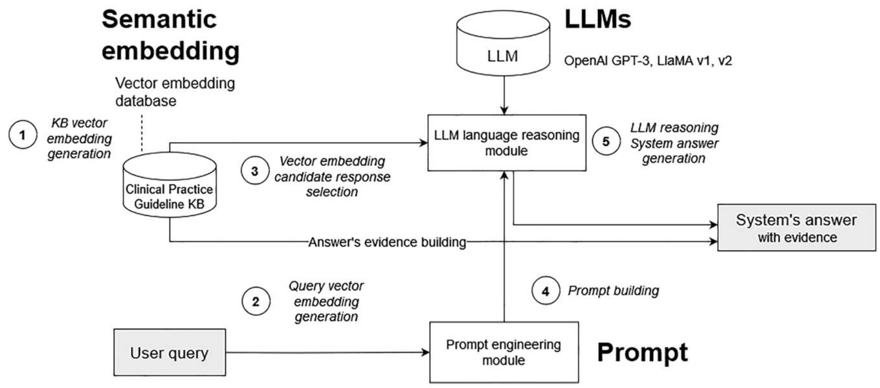

Figure 1. Architecture proposed.

replies ordered by cosine distance to the query vector. The number of candidates used varies depending on the context model size.

The prompt engineering module (step 4) integrates all components: we ask the question and provide the model with the context that it should consider to generate the answer. The prompting forces the model to formulate answers using only the reliable contents. An important part of the prompt is the directive preventing the model from inventing the answer.

Next, we provide a sample prompt:

Respond to the question based on the following context. If the question cannot be answered based on the context, say 'I do not have information to answer this question'.
Context:
Candidate paragraph 1
Candidate paragraph 2
...
Question: *For a person with ADHD, what is the effectiveness and safety of training and education programmes aimed at parents, mothers, family members, caregivers, and other reference persons in their environment?*
Answer:

On step 5 (LLM reasoning), the model reasons on the information provided whether it is possible to formulate an answer. This computation is performed by the LLM engine. The LLM determines whether a suitable answer is present among the paragraphs chosen based on their proximity in the embeddings space. If not, it responds that it lacks sufficient information to answer the user's question. If there is information in the context, the system answer also includes the scientific evidence in which the response was based on.

Therefore the system we developed and evaluated does not use LLMs as a knowledge base. Instead, the knowledge for the answers it provides originates directly from the Clinical Practice Guidelines (CPGs). Specifically, the system either responds with exact fragments from the CPGs or indicates its lack of knowledge on the subject, eliminating any chance of fabricated ('hallucinated') responses.

The LLM's role is limited to two key steps in processing queries:

(1) Upon receiving a query, the system searches for the most relevant paragraphs within the space of paragraph embeddings.
(2) The LLM then analyses these candidate paragraphs to select an answer derived directly from the CPGs, ensuring the accuracy and traceability of its content.

4. Materials and methods

In this section we describe how we have used our proposal to generate Q&A systems that respond to questions about depression and attention deficit hyperactivity disorder (ADHD).

4.1. Clinical practice guidelines

The Aragonese Institute of Health Sciences (IACS)[3] is a crucial part of the Spanish National Health System. One of its primary responsibilities is the production of Clinical Practice Guidelines (CPGs). These guidelines are developed by a team of healthcare experts, and their purpose is to provide medical professionals with recommendations for patient care that are founded on solid scientific research.

To ensure that these guidelines remain updated with the latest advancements in science and technology, the IACS makes it a regular practice to review and revise the CPGs. By doing so, they guarantee that the information is up-to-date and reflects the most recent understanding in the medical field.

The IACS has an extensive catalogue of clinical guidelines,[4] containing several dozens of medical guidelines. These guidelines are written in the Spanish language and are available either in HTML or PDF format. Though the specific content of each guide may vary, they generally consist of sections that cover various aspects of medical care. These include definitions of medical conditions, diagnostic procedures, risk factors associated with different diseases, methods for evaluating and screening patients, first-hand accounts of patient experiences, detailed information on treatments, and strategic approaches to treatment.

In creating these guidelines, the authors employ evidence levels that have been adapted from the 'Scottish Intercollegiate Guidelines Network' handbook (SIGN 2014). This ensures a standardised approach to evaluating and presenting evidence.

The CPGs' catalogue was created to preserve the specialised knowledge contained within these guides. Nonetheless, the information presented may be challenging to grasp and interpret, and it is also difficult to locate answers to specific questions without revising a comprehensive part of the text. We recognise the importance of making this information more accessible and quearible and with this objective in mind, we have embarked on a project to enhance the overall system's accessibility.

As part of this effort, a proof of concept has been built for two distinct medical domain guides related specifically to mental health: clinical practice guideline on the management of depression in adults[5] and clinical practice guideline on therapeutic interventions in ADHD.[6] ⑤成人抑鬱症 ⑥多動症

Those guidelines are directed to a wide range of professionals participating in the treatment and outpatient care for those mental issues and can also be useful for patients themselves, their families and caregivers.

The guide for depression includes definition, risk factors, suicide risk, diagnosis, perspectives and experiences of patients with depression and their families, evaluation and screening for depression, models of depression care, general principles of management, collaborative care (including Cognitive-Behavioral Therapies, Behavioral Activation, Problem-Solving Therapy and Counselling among others), pharmacological treatment, strategies in treatment of resistant depression, diagnostic and therapeutic strategies, quality indicators, to mention the main topics.

With regard to ADHD, a special emphasis is placed on educational adaptation and social health. The guide covers topics such as: definitions, diagnosis, academic education, social skills, and integration into the job market, pharmacological interventions, therapeutic interventions and combinations, interventions targeting the environment of individuals with ADHD, among other aspects of ADHD.

Our innovative approach allows the system to respond to questions about the contents of the guides, locate relevant information in the guides, reason based on content that has been validated by experts, and clearly demonstrate the evidence that underpins the information. This represents a significant step towards making specialised mental health knowledge more readily available to those who need it.

4.2. Question answering systems for depression and ADHD

First, we preprocessed the CPGs. The typical CPG web layout is a HTML page where the table of contents links to different sections. We divided them into paragraphs, with each one serving as an individual piece of information, and subsequently converted every paragraph into an embedding representation as explained in Section 3.

Then, we used the approach depicted in Figure 1 to generate Q&A systems for depression and ADHD. Each of the two available in three variations, depending on the kind of underlying generative language models that implement the reasoning: GPT-3, LlaMA or LlaMA-2. We have used both GPT and LlaMA for several reasons. On the one hand, OpenAI has created the most powerful state of the art for most NLP tasks (Brown et al. 2020), and this is why we use OpenAI GPT-3. However, using a proprietary system entails several disadvantages:

- The exploitation costs are higher.
- The queries of specialists and users to the system are registered by the company. In fact, several authors have expressed ethical concerns related to the collection of the data inputted into LLMs and the lack of transparency about how it is stored and used (Li et al. 2023).
- If this infrastructure supports a critical citizen service, the Public Administration would have no control over its availability and quality of service.
- It is not possible to audit the models to assess, for example, whether they have any kind of bias or discriminatory behaviour.
- The behaviour of the system must be repeatable over time.

Hence, relying on a corporate-owned model may pose long-term challenges, such as limited control over its management. Additionally, there are concerns regarding the security of the data exchanged with the model. In order to address these questions we also have used open-source LLMs such as LlaMA and LlaMA-2.

Therefore, we have produced our systems with the three following models:

(3) A foundational generative language model GPT-3 developed by OpenAI (Brown et al. 2020). This model is trained on proprietary datasets, and the details of the architecture and training are confidential. For our proof of concept we used the free API access to this model `text-davinci-003` available at OpenaAI site.[7]

(4) LlaMA, a family of open-source high-performing foundational language models (Touvron, Lavril, et al. 2023) ranging from 7 B to 65 B parameters. LLaMA is restricted to non-commercial use and is not available for production. However, it demonstrates that even when the specific knowledge contained within the model is not required, the reasoning, logical and semantic capabilities of the model are preserved. This holds true even if it is not trained on proprietary datasets. We use Nous-Hermes-13b, a state-of-the-art LLaMA model fine-tuned on over 300,000 instructions (NousResearch 2023).

(5) LlaMA-2 or LLaMA 2-Chat (Touvron, Martin, et al. 2023), a family of generative models based on LlaMA and fine-tuned specifically for dialogue use cases. The model and weights are licensed for both researchers and commercial entities and therefore can be used in the commercial production. For this purpose, we also use a Nous-Hermes finetuned version.

商業授權

LLMs are rapidly advancing, however, the computational resources required to deploy these systems, such as GPUs with substantial memory, are significant. A positive aspect of our proposal is that it can be used with different LLMs, including lightweight and open models, and thus can be adopted by stakeholders with a varied range of capabilities and requirements. Therefore, we fond it worthwhile to examine smaller and medium-sized language models that can operate on scalable middleware. Consequently, we have explored in our experimental setup both OpenAI models and light and open models, such as LlaMA and LlaMA-2, that do not require extensive contexts and offer more cost-effective and privacy-conscious use for interested parties.

在可擴展中間件上運行

Recently, OpenAI has released capabilities for GPT-4-Turbo to process PDF documents to generate chatbots for downstream tasks. However, GPT-4, though a powerful model, appears less suitable for our objectives. This is particularly because the CPGs in PDF format that we use cannot be fully processed by GPT-4-Turbo models, as they exceed their 128 K token context limit.

Regarding the multilingual capabilities of the models it is important to note that they have undergone training using comprehensive multilingual datasets. Current assessments, as highlighted in Lai et al. (2023), demonstrate their proficiency in most high-resource languages other than English. This gives us confidence in their ability to process and reason using a Spanish knowledge database. Moreover, specific studies, such as those examining ChatGPT, indicate its capability to directly analyse and reason at Spanish medical exams (Madrid-García et al. 2023).

5. Experiments

In order to evaluate the quality of the answers generated by the Q&A systems developed, we designed the evaluation setup described in Section 5.1. The results achieved are presented and discussed in Section 5.2.

5.1. Evaluation setup

To assess the models, we recruited an expert in both depression and ADHD and asked him to formulate 10 questions for each scenario and their corresponding correct answers, which are reported in Appendix 1 (translated into English from the original in Spanish). Then, we submitted these questions to each question-answering system (depression and ADHD with different models). The questions pertain to the information within the relevant Clinical Practice Guidelines (CPGs), and the large language models (LLMs) underlying the question-answering systems remain unaffected by these questions.

The expert scored the model outputs according to the quality metrics defined, which are listed below. In line with the research on the evaluation of medical question-answering using LLMs (Guo et al. 2023; Xie et al. 2023; Ziyu et al. 2023), we propose the combination of the three semantic qualities of the answer as an evaluation criteria:

(3) **Coherence** encompasses both the degree to which it is related to the corresponding question, and also clear and linguistically well articulated.

(4) **Veracity** stands as our top priority for system responses, especially when conveying mental health knowledge where information accuracy is critical.

(5) **Evidence** focuses on whether the system was able to correctly provide accurate scientific basis for the response.

Table 2. Metrics used for evaluating the models' answers.

Metric description	Evaluation scale
Coherence: The answer generated by the system is logically structured and clear (opposite of vague), has all the necessary information (complete), does not contain irrelevant information, and does not include contradictions.	[ ] 1–Strongly disagree
	[ ] 2–Disagree
	[ ] 3–Neither agree or disagree
	[ ] 4–Agree
	[ ] 5–Strongly agree
Veracity: The answer generated by the system is correct and relevant to the question.	[ ]−1–The answer is incorrect or it is 'I do not have the information to answer that question'
	[ ] 0–The answer is correct but not relevant.
	[ ] −1–The answer is correct
Evidence: The answer generated by the system includes references to relevant studies, clinical trials, or guidelines that can validate it.	[ ] 1–Strongly disagree
	[ ] 2–Disagree
	[ ] 3–Neither agree or disagree
	[ ] 4–Agree
	[ ] 5–Strongly agree

Table 2 shows the metrics described, to rate them we propose to use Likert scales, in a similar way as it is done with similar evaluation criteria in the literature (Balel 2023; Chen et al. 2023; Rasmussen et al. 2023; Xie et al. 2023).

We would like to highlight that we are not evaluating answers invented by a large language model. Instead, we assess the ability of the question-answering system to reason over the question using the LLMs and finding and composing a trusted response from the appropriate pieces of information present in the validated knowledge base (the clinical guides).

If the system response to a question is that it cannot provide an answer, the medical evaluator has determined that it should be marked as incorrect, receiving the lowest possible score on both coherence and evidence scales. Thus, we have adopted a strict policy in which not answering is penalised, even though that is the best possible answer a trustfully system can give without inventing a response.

A second medical expert confirmed the accuracy of the assessments from a wider medical perspective. During the peer review process, all assessments of the first expert were considered free from any errors by the second expert.

The assessment of various questions, model responses, along with the expert's evaluations and corresponding correct answers, can be found in Appendix 2.

5.2. Discussion of results

Table 3 presents the results of the depression question-answering system. For each metric (coherence, veracity, and evidence) we present the results of the three LLMs employed (GPT-3, LlaMA-1 and LlaMA-2). The first column represents the possible values on the Likert scale for each metric (see Table 2), each cell corresponds to the number of times that option has been selected per model. The cell values in the last row (Total) are calculated as the scalar product of the number of answers of each type by the corresponding Likert scale value.

The results show that LlaMA-1 outperforms both OpenAI's GPT-3 and LlaMA-2 across all metrics. Due to our system penalising for not providing an answer, LlaMA-1 successfully generated responses to expert questions in all but one instance, while GPT-3 failed to answer three times. Regarding the quality of the answers, both LlaMA-1 and LlaMA-2 demonstrated similar performance in terms of answer coherence and the presence of supporting evidence.

Table 4 presents the results for the ADHD domain. In this area, GPT-3's performance is notably weak, failing to answer expert questions in 25% of cases. The open-source models, LlaMA-1 and LlaMA-2, exhibit

Table 3. Results for depression.

Coherence	GPT-3	LLaMA-1	LLaMA-2	Total
1	3	1	2	6
2	1	0	0	1
3	1	1	1	3
4	10	6	5	21
5	9	16	16	41
Total	93	**108**	105	
Veracity	GPT-3	LLaMA-1	LLaMA-2	Total
1	3	1	2	6
2	0	0	0	0
3	21	23	22	66
Total	18	**22**	20	
Evidence	GPT-3	LLaMA-1	LLaMA-v2	Total
1	3	1	2	6
2	0	0	0	0
3	1	2	0	3
4	15	18	15	48
5	5	3	6	14
Total	91	**94**	92	

Table 4. Results for ADHD.

Coherence	GPT-3	LlaMA-1	LLaMA-2	Total
1	6	3	3	12
2	2	0	1	3
3	0	0	1	1
4	8	9	3	20
5	8	12	16	36
Total	82	99	**100**	
Veracity	GPT-3	LLaMA-1	LLaMA-2	Total
1	6	3	3	12
2	1	0	0	1
3	17	21	21	59
Total	11	**18**	**18**	
Evidence	GPT-3	LLaMA-1	LLaMA-v2	Total
1	6	3	3	12
2	1	0	0	1
3	1	0	0	1
4	13	11	11	35
5	3	10	10	23
Total	78	**97**	**97**	

nearly identical effectiveness, with LlaMA-2 showing a slight edge in the coherence of its answers.

Analysing the combined results across all models and domains, it is evident that the open-source models, LlaMA-1 and LlaMA-2, respond more frequently than the proprietary GPT-3. This results in their superior performance in both ADHD and depression fields across all metrics. Overall, the performance comparison of LlaMA-1 and LlaMA-2 indicates similar outcomes, suggesting that the linguistic capabilities of these models do not significantly increase with their size (Table 5).

In global terms, we can conclude that our proposal shows positive results for answer coherence, veracity and the provision of evidence. Regarding the comparison of the models used, open-source models, LlaMA-1 nd LlaMA-2, obtain the best results overall. It should be noted that these models are open and their use may have certain advantages when it comes to cost-efficiency, control of the model, availability and quality of service, privacy of the queries, and other relevant issues previously highlighted.

6. Limitations

Despite the positive results, our proposal has several limitations. On the one hand, language models still have a limited content window used for question formulation. This fact restrict the amount of truthful content we are able to include in the context for the system to reason and use as a candidate reply. Although this restriction has not affected the quality of the responses in our application domains, it could be a factor to be considered in other scenarios. In addition, in the proposed architecture, there are limitations in the search space for candidate answers to solve the question posed by the user. If this type of system is to be used on very large knowledge bases, it will be necessary to develop some kind of hierarchical indexing system to partition search space and additionally to improve the embedding representation in the clinical context.

Another constraint of our proof-of-concept system is its assessment by just one expert specialised in our target domains. A second expert provided a broader oversight of the evaluation process.

Lastly, the support of different languages is very disparate. Although there is the possibility of translating the content, the same level of quality of the system's output is not guaranteed in all languages. The results described have been produced for Spanish.

7. Security and privacy

The target audience for the question-answering systems developed are healthcare professionals, who can utilise these systems for efficient access to extensive clinical practice guidelines. Our proposal prevents the system from inventing responses. The remaining concern in the possibility that the system provides irrelevant information in response to a query. Such information would in any case be retrieved from the trusted guides, as we assure that the system will not create or deliver information outside the established knowledge base. In such situations, medical professionals can use their expertise to discern discrepancies between the query and the response. In addition, the fact that the system provides evidence and traceability of the sources of the response, facilitates that the users can verify the response against the guidelines.

Table 5. Aggregated results.

Guide	Models	Coherence	%	Veracity	%	Evidence	%
Depression	GPT-3	93	78%	18	75%	91	76%
	LlaMA v1	108	90%	22	92%	94	78%
	LlaMA v2	105	88%	20	83%	92	77%
ADHD	GPT-3	82	68%	11	46%	78	65%
	LlaMA v1	99	83%	18	75%	97	81%
	LlaMA v2	100	83%	18	75%	97	81%
TOTAL	GPT-3	175	73%	29	60%	169	70%
	LlaMA v1	207	86%	40	83%	191	80%
	LlaMA v2	205	85%	38	79%	189	79%

It must be taken into account that the objective of the system is to query the clinical practice guidelines, which are about common medical knowledge rather than specific patient details. Therefore, it is unlikely that queries will contain personal patient information. Nevertheless, the users of the system, who are medical professionals accustomed to adhering to GDPR regulations, are advised to refrain from submitting patient-specific information in their queries as personal information is not required for the system's functionality.

The fact that our proposal obtains very positive results with open models (see Section 5.2), makes it possible for adopters to consider on-premise options that do not require sending the user queries to external servers, thus facilitating security and privacy control.

Furthermore, the content of the guides is public and devoid of personal or sensitive information, minimising the risk of bias in the system's responses, which are based solely on this verified content.

8. Conclusions and future work

We have proposed a method to develop Q&A systems based on off-the-shelve large language models in the mental health domain. Our approach exploits the well-known capacities of the LLMs for logical and semantic reasoning while circumscribing the generated responses to reliable contents of clinical practice guidelines.

As a proof-of-concept, we have followed our proposal to develop several Q&A systems which include a combination of proprietary (OpenAI's GPT-3) and open (LlaMA-1 and LlaMA-2) LLMs and two mental health scenarios: depression and ADHD.

We have evaluated the systems using three metrics related to (a) coherence of the answer, (b) whether it is correct or not, and (c) presence of scientific evidence that underpins it. Our first medical expert, specialised in our areas of focus, assessed the models across all metrics within each domain. A second medical expert then corroborated these evaluations based on general medical knowledge.

The empirical results show a high level of truthfulness in the answer, i.e. the contents of the response are accurate; high coherence in the exposition of the answer (with regard to its relation to the question posed); and the desired ability to cite the evidence (excerpts of the clinical guide) on which the answer is based. In addition, we have experimented with proprietary and open models, obtaining good quality in both types of language models. The result with open models is particularly promising because of their applicability for mental health applications with high privacy constraints.

Thus, our proposal is a first step in the direction of using LLMs more safely to provide high quality reliable responses in the medical and mental health domain. Expert training and involvement are two key aspects identified in the literature to embrace LLM models more safely in medical applications (Karabacak and Margetis 2023). The fact that the Q&A systems developed are mainly directed to experts, also facilitate that professionals understand the capabilities and limitations of the technology and explore it safely to incorporate it in their clinical practice; at the same time experts have been involved in providing valuable feedback on the usefulness and usability of the LLM-generated responses.

Despite the positive results achieved, there are still numerous challenges to be addressed in mental health Q&A systems:

- Medical terminology and contextual understanding: mental health involves a vast and specialised vocabulary, including medical terms, acronyms, and jargon.
- Effective capacity to integrate new sources of information and to update existing content. Mental health knowledge is continuously evolving, with new research findings, treatments, and guidelines emerging regularly.
- Privacy and Security: mental health data is highly sensitive and subject to strict privacy regulations. Ensuring that question-answering systems handle and store data securely is a paramount concern.
- Multilingual support. Multilingual support of the major language models is very disparate. Good language models in all languages or an acceptable translation of the system's questions and answers is necessary.
- Multiple communication channels (web, voice, message systems, etc.). Adaptation of the answers and the language of consultations is necessary depending on the communication channel used.
- Allow multiple linguistic registers; from the most technical levels, used by clinicians, to simplified language, used by the non-specialised citizen. Mental health questions often contain ambiguous or vague language, and the same term can have different meanings in different contexts.

Finally, for future work we will focus on the following elements:

- Experimentation with other LLMs that guarantee the highest quality of responses but allow lower technical requirements for the operation of the system.
- Analysis of the multilingual support of the system.

- Evaluation of the integrated use of text simplifiers to allow the use of the system by non-specialists.
- Analysis of the system on wider knowledge bases.
- Inclusion of LLM engine ensembles. When the knowledge base is very large, it is possible to specialise the models so that their reasoning is better adapted to specific domains. In this way, the system would pre-select the most appropriate engine to manage the response to the user.
- Inclusion of a hierarchical semantic search and indexing subsystem. It could be useful when there is a very large knowledge base to be consulted simultaneously. Hierarchical organisation of information by means of hierarchical semantic indexes facilitates the search for information in large knowledge bases.
- Extensions of the architecture for multi-modal multi-channel systems that require specific adaptation of the content offered in the response.

Notes

1. https://crfm.stanford.edu/2023/03/13/alpaca.html
2. https://platform.openai.com/docs/guides/embeddings/what-are-embeddings
3. https://www.iacs.es/
4. https://portal.guiasalud.es/
5. https://portal.guiasalud.es/egpc/depresion-adulto-introduccion/
6. https://portal.guiasalud.es/egpc/tdah-introduccion/
7. https://platform.openai.com/docs/models/gpt-3

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This publication is part of the 'CONVERSA: Effective and efficient resources and models for transformative conversational AI in Spanish and co-official languages' project with reference (Agencia Estatal de Investigación) TED2021-132470B-I00, funded by MCIN/AEI/10.13039/501100011033 and by the European Union 'NextGenerationEU/PRTR' and the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR, https://menhir-project.eu).

Funding

This publication is part of the 'CONVERSA: Effective and efficient resources and models for transformative conversational AI in Spanish and co-official languages' project with reference (Agencia Estatal de Investigación) TED2021-132470B-I00, funded by MCIN/AEI/10.13039/501100011033 and by the European Union 'NextGenerationEU/PRTR' and the European Union's Horizon 2020 research and innovation programme under grant agreement No 823907 (MENHIR, https://menhir-project.eu).

References

Antaki, F., S. Touma, D. Milad, J. El-Khoury, and R. Duval. 2023. "Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of Its Successes and Shortcomings." *Ophthalmology Science* 3 (4): 100324. https://doi.org/10.1016/j.xops.2023.100324.

Arias, D., S. Saxena, and S. Verguet. 2016. "Extracting PICO Sentences From Clinical Trial Reports Using Supervised Distant Supervision." *EClinicalMedicine* 54:101675. https://doi.org/10.1016/j.eclinm.2022.101675.

Arif, T. B., U. Munaf, and I. Ul-Haque. 2023. "The Future of Medical Education and Research: Is ChatGPT a Blessing Or Blight in Disguise?" *Medical Education Online* 28 (1): 2181052. https://doi.org/10.1080/10872981.2023.2181052.

Ayers, J. W., A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, et al. 2023. "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum." *JAMA Internal Medicine* 183. https://doi.org/10.1001/jamainternmed.2023.1838.

Balel, Y. 2023. "Can ChatGPT Be Used in Oral and Maxillofacial Surgery?" *Journal of Stomatology, Oral and Maxillofacial Surgery* 101471. https://doi.org/10.1016/j.jormas.2023.101471.

Benoit, J. R. A. 2023. "ChatGPT for Clinical Vignette Generation, Revision, and Evaluation." medRxiv. https://doi.org/10.1101/2023.02.04.23285478.

Bowman, S. R. 2023. "Eight Things to Know About Large Language Models." arXiv 2304.00612.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, Vol. 33, 1877–1901. Red Hook, NY, USA: Curran Associates Inc.

Cabrera, J., M. S. Loyola, I. Magaña, and R. Rojas. 2023. "Ethical Dilemmas, Mental Health, Artificial Intelligence, and LLM-Based Chatbots." In *Bioinformatics and Biomedical Engineering*, 313–326. Berlin, Germany: Springer-Verlag.

Callejas, Z., and D. Griol. 2021. "Conversational Agents for Mental Health and Wellbeing." In *Dialog Systems: A Perspective From Language, Logic and Computation*, 219–244. Heidelberg, Germany: Springer International Publishing. https://doi.org/10.1007/978-3-030-61438-6_11.

Chen, S., M. Wu, K. Q. Zhu, K. Lan, Z. Zhang, and L. Cui. 2023. "LLM-Empowered Chatbots for Psychiatrist and Patient Simulation: Application and Evaluation." arXiv 2305.13614.

Chronopoulou, A., C. Baziotis, and A. Potamianos. 2019. "An Embarrassingly Simple Approach for Transfer Learning From Pretrained Language Models." In *Proc. of NAACL-HLT'19*, 2089–2095. Minneapolis, Minnesota. https://doi.org/10.18653/v1/N19-1213.

Corrigan, P., and A. Matthews. 2016. "Stigma and Disclosure: Implications for Coming out of the Closet." *Journal of Mental Health* 12 (3): 235–248. https://doi.org/10.1080/0963823031000118221.
Dahmen, J., M. E. Kayaalp, M. Ollivier, A. Pareek, M. T. Hirschmann, J. Karlsson, and P. W. Winkler. 2023. "Artificial Intelligence Bot ChatGPT in Medical Research: The Potential Game Changer As a Double-edged Sword." *Knee Surgery, Sports Traumatology, Arthroscopy* 31 (4): 1187–1189. https://doi.org/10.1007/s00167-023-07355-6.
Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." ArXiv abs/1810.04805.
DiGiorgio, A. M., and J. M. Ehrenfeld. 2023. "Artificial Intelligence in Medicine & ChatGPT: De-Tether the Physician." *Journal of Medical Systems* 47 (1). https://doi.org/10.1007/s10916-023-01926-3.
Duong, D., and B. Solomon. 2023. "Analysis of Large-language Model Versus Human Performance for Genetics Questions." *European Journal of Human Genetics*. https://www.nature.com/articles/s41431-023-01396-8#citeas.
Eysenbach, G. March 6, 2023. "The Role of ChatGPT, Generative Language Models, and Artificial Intelligence in Medical Education: A Conversation With ChatGPT and a Call for Papers." *JMIR Medical Education* 9:e46885. https://doi.org/10.2196/46885.
Gao, Y., Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, and H. Wang. 2024. "Retrieval-Augmented Generation for Large Language Models: A Survey".
Garg, R. K., V. L. Urs, A. A. Agrawal, S. K. Chaudhary, V. Paliwal, and S. K. Kar. 2023. "Exploring the Role of ChatGPT in Patient Care (Diagnosis and Treatment) and Medical Research: A Systematic Review." medRxiv. https://doi.org/10.1101/2023.06.13.23291311.
Ge, Y., Y. Guo, S. Das, M. A. Al-Garadi, and A. Sarker. 2023. "Few-shot Learning for Medical Text: A Review of Advances, Trends, and Opportunities." *Journal of Biomedical Informatics* 144:104458. https://doi.org/10.1016/j.jbi.2023.104458.
Gu, Y., R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. 2021. "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing." *ACM Transactions on Computing for Healthcare* 3 (1). https://doi.org/10.1145/3458754.
Guo, Z., R. Jin, C. Liu, Y. Huang, D. Shi, L. Yu, Y. Liu, et al. 2023. "Evaluating Large Language Models: A comprehensive Survey".
Haidich, A. 2010. "Meta-analysis in Medical Research." *Hippokratia* 14:29–37.
Harrer, S. 2023. "Attention is Not All You Need: The Complicated Case of Ethically Using Large Language Models in Healthcare and Medicine." *EBioMedicine* 90:104512. https://doi.org/10.1016/j.ebiom.2023.104512.
Ji, S., T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria. 2021. "MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare." arXiv 2110.15621.
Kaplan, J., S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, et al. 2020. "Scaling Laws for Neural Language Models." arXiv 2001.08361.
Karabacak, M., and K. Margetis. 2023. "Embracing Large Language Models for Medical Applications: Opportunities and Challenges." *Cureus* 15 (5): e39305. https://doi.org/10.7759/cureus.39305.
Kononenko, I. 2001. "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective." *Artificial Intelligence in Medicine* 23 (1): 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X.
Lai, V. D., N. T. Ngo, A. P. B. Veyseh, H. Man, F. Dernoncourt, T. Bui, and T. H. Nguyen. 2023. "Chatgpt Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning".
Lamichhane, B. 2023. "Evaluation of ChatGPT for NLP-based Mental Health Applications." arXiv 2303.15727.
Lee, E. W., B. C. Wallace, K. I. Galaviz, and J. C. Ho. 2020. "MMiDaS-AE: Multi-Modal Missing Data Aware Stacked Autoencoder for Biomedical Abstract Screening." In *Proc. of CHIL'20*, 139–150. Toronto, Ontario, Canada. https://doi.org/10.1145/3368555.3384463.
Levine, D. M., R. Tuwani, B. Kompa, A. Varma, S. G. Finlayson, A. Mehrotra, and A. Beam. 2023. "The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model." medRxiv. https://doi.org/10.1101/2023.01.30.23285067.
Li, H., J. T. Moon, S. Purkayastha, L. A. Celi, H. Trivedi, and J. W. Gichoya. 2023. "Ethics of Large Language Models in Medicine and Medical Research." *The Lancet Digital Health* 5 (6): e333–e335. https://doi.org/10.1016/S2589-7500(23)00083-3.
Liu, S., A. P. Wright, B. L. Patterson, J. P. Wanderer, R. W. Turer, S. D. Nelson, A. B. McCoy, D. F. Sittig, and A. Wright. 2023a. "Assessing the Value of ChatGPT for Clinical Decision Support Optimization." medRxiv. https://doi.org/10.1101/2023.02.21.23286254.
Liu, S., A. P. Wright, B. L. Patterson, J. P. Wanderer, R. W. Turer, S. D. Nelson, A. B. McCoy, D. F. Sittig, and A. Wright. 2023b. "Using AI-generated Suggestions From ChatGPT to Optimize Clinical Decision Support." *Journal of the American Medical Informatics Association* 30 (7): 1237–1245. https://doi.org/10.1093/jamia/ocad072.
Liu, X., Y. Zhou, and Z. Wang. 2019. "Recognition and Extraction of Named Entities in Online Medical Diagnosis Data Based on a Deep Neural Network." *Journal of Visual Communication and Image Representation* 60:1–15. https://doi.org/10.1016/j.jvcir.2019.02.001.
Madrid-García, A., Z. Rosales-Rosado, D. Freites-Nuñez, I. Pérez-Sancristobal, E. Pato-Cour, C. Plasencia-Rodríguez, L. Cabeza-Osorio, et al. 2023. "Harnessing ChatGPT and GPT-4 for Evaluating the Rheumatology Questions of the Spanish Access Exam to Specialized Medical Training." medRxiv. https://doi.org/10.1101/2023.07.21.23292821.
McCullock, S. P., and R. M. Scrivano. 2023. "The Effectiveness of Mental Illness Stigma-reduction Interventions: A Systematic Meta-review of Meta-analyses." *Clinical Psychology Review* 100:102242. https://doi.org/10.1016/j.cpr.2022.102242.
Miwa, M., J. Thomas, A. O'Mara-Eves, and S. Ananiadou. 2014. "Reducing Systematic Review Workload Through Certainty-based Screening." *Journal of Biomedical Informatics* 51:242–253. https://doi.org/10.1016/j.jbi.2014.06.005.
Murad, M. H., N. Asi, M. Alsawas, and F. Alahdab. 2016. "New Evidence Pyramid." *BMJ Evidence-Based Medicine* 21 (4): 125–127. https://doi.org/10.1136/ebmed-2016-110401.

Nori, H., N. King, S. M. McKinney, D. Carignan, and E. Horvitz. 2023. "Capabilities of GPT-4 on Medical Challenge Problems." arXiv 2303.13375.

NousResearch. 2023. "Nous-hermes-13b." Hugging Face. Accessed August 8, 2023. https://huggingface.co/NousResearch/Nous-Hermes-13b.

Patel, S. B., and K. Lam. 2023. "ChatGPT: the Future of Discharge Summaries?" *The Lancet Digital Health* 5 (3): e107–e108. https://doi.org/10.1016/S2589-7500(23)00021-3.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." https://api.semanticscholar.org/CorpusID:160025533.

Rasmussen, M. L. R., A. C. Larsen, Y. Subhi, and I. Potapenko. 2023. "Artificial Intelligence-based ChatGPT Chatbot Responses for Patient and Parent Questions on Vernal Keratoconjunctivitis." *Graefe's Archive for Clinical and Experimental Ophthalmology*. https://doi.org/10.1007/s00417-023-06078-1.

Richens, J., C. Lee, and S. Johri. 2020. "Improving the Accuracy of Medical Diagnosis with Causal Machine Learning." *Nature Communications* 11:3923. https://doi.org/10.1038/s41467-020-17419-7.

Sallam, M. 2023. "The Utility of ChatGPT as an Example of Large Language Models in Healthcare Education, Research and Practice: Systematic Review on the Future Perspectives and Potential Limitations." medRxiv. https://doi.org/10.1101/2023.02.19.23286155.

Scerri, A., and K. H. Morin. 2023. "Using Chatbots Like ChatGPT to Support Nursing Practice." *Journal of Clinical Nursing* 32 (15–16): 4211–4213. https://doi.org/10.1111/jocn.16677.

SIGN. 2014. *Scottish Intercollegiate Guidelines Network- SIGN 50: A Guideline Developer's Handbook*. Healthcare Improvement Scotland.

Singhal, K., S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, et al. 2023. "Large Language Models Encode Clinical Knowledge." *Nature* 620 (7972): 172–180. https://doi.org/10.1038/s41586-023-06291-2.

Singhal, K., T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, et al. 2023. "Towards Expert-Level Medical Question Answering with Large Language Models." arXiv 2305.09617.

Taylor, R., M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic. 2022. "Galactica: A Large Language Model for Science." arXiv 2211.09085.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M. A. Lachaux, T. Lacroix, B. Rozière, et al. 2023. "LLaMA: Open and Efficient Foundation Language Models." ArXiv abs/2302.13971.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, et al. 2023. "Llama 2: Open Foundation and Fine-Tuned Chat Models." ArXiv abs/2307.09288.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. "Attention is All You Need." In *Advances in Neural Information Processing Systems*, edited by I. Guyon et al., Vol. 30. New York, USA: Curran Associates, Inc.

Wallace, B. C., J. Kuiper, A. Sharma, M. Zhu, and I. J. Marshall. 2016. "Extracting PICO Sentences From Clinical Trial Reports Using Supervised Distant Supervision." *Journal of Machine Learning Research* 17. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5065023/.

Xie, Y., I. Seth, W. M. Rozen, and D. J. Hunter-Smith. 2023. "Evaluation of the Artificial Intelligence Chatbot on Breast Reconstruction and Its Efficacy in Surgical Research: A Case Study." *Aesthetic Plastic Surgery*. https://doi.org/10.1007/s00266-023-03443-7.

Yang, R., T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu. 2023. "Large Language Models in Health Care: Development, Applications, and Challenges." *Health Care Science*. https://doi.org/10.1002/hcs2.61.

Yousaf, F., S. Iqbal, N. Fatima, T. Kousar, and M. S. M. Rahim. 2023. "Multi-class Disease Detection Using Deep Learning and Human Brain Medical Imaging." *Biomedical Signal Processing and Control* 85:104875. https://doi.org/10.1016/j.bspc.2023.104875.

Ziyu, Z., C. Qiguang, M. Longxuan, L. Mingda, H. Yi, Q. Yushan, B. Haopeng, et al. August 2023. "Through the Lens of Core Competency: Survey on Evaluation of Large Language Models." In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Vol. 2: Frontier Forum)*, edited by J. Zhang, 88–109. Harbin, China: Chinese Information Processing Society of China. https://aclanthology.org/2023.ccl-2.8.

Appendices

Appendix 1. Questions for the guidelines (originally in Spanish)

A.1. Clinical practice guideline on the management of depression in adults

(3) What are the perspectives of patients and relatives regarding depression and their experiences with the healthcare they received?
(4) How should the evaluation for depression be conducted?
(5) Which scales have the best psychometric properties for the evaluation of depression in adults?
(6) What is the efficacy and safety of antidepressant drugs in the treatment of depressive episodes in adults?
(7) How long and at what dosage should pharmacological treatment be maintained after the remission of the depressive episode?
(8) Which pharmacological strategies are most effective in patients with treatment-resistant depression?
(9) What is the efficacy and safety of electroconvulsive therapy as a treatment for depression?
(10) What is the efficacy and safety of vagus nerve stimulation as an adjunctive treatment for treatment-resistant depression?
(11) What is the efficacy and safety of transcranial magnetic stimulation as an adjunctive treatment for treatment-resistant depression?
(12) What is the role of psychotherapy as an enhancement or alternative in patients with treatment-resistant depression?
(13) Does screening improve health outcomes in depression?
(14) What is the effectiveness of stepped care and collaborative care models?
(15) What is the effectiveness of different psychological interventions in patients with depression?
(16) Is physical exercise effective in patients with depression?
(17) What is the efficacy and safety of St. John's Wort in the treatment of adult depression?
(18) What are the indicators that allow for monitoring the quality in the management of depression?
(19) What is the development and impact of decision support and knowledge management systems in the management of depression?
(20) What are the risk factors for depression?
(21) What should be primarily considered in the assessment of suicide risk in patients with depression?
(22) What is the stepped care model in the treatment of depression?
(23) What are the main psychotherapeutic interventions in the treatment of depression?
(24) What are the pharmacokinetic interactions of antidepressants?
(25) What are the contraindications of Bupropion?
(26) Are benzodiazepines recommended for the treatment of depression?

A.2. Clinical practice guideline on therapeutic interventions in attention deficit hyperactivity disorder

(3) Throughout the formative stages, in people with ADHD, what guidelines are appropriate to improve organisation and time management dedicated to school tasks and academic performance?
(4) During the educational stages of people with ADHD, and regardless of the type or combination of treatments, what kind of procedure or coordination model between the academic institution, the health system, and the family environment improves educational outcomes?
(5) Are training and social skills training interventions effective and safe for people with ADHD, regardless of any associated comorbidities they may have, to integrate these individuals into the labor market, or, once integrated, to allow them to maintain appropriate relationships in the work environment?
(6) In the context of Secondary Education, which interventions should primarily be emphasised for students with ADHD?
(7) To facilitate the integration of people with ADHD into the labor market or their retention within it, what cognitive-behavioural programmes can be used?
(8) What are the reasons why school/educational interventions should be included in the treatment programme for people with ADHD?
(9) What exception is there to the use of drugs for ADHD in children under 6 years old?
(10) What medications are indicated for the treatment of ADHD in girls and boys?
(11) What do systemic therapies for the treatment of ADHD consist of?
(12) What is the goal of psychoanalytic therapies in the treatment of ADHD?
(13) In people with ADHD, what is the effectiveness and safety of school/educational intervention programmes?
(14) The indication for pharmacological treatment in people with ADHD, to what extent is it effective and safe?
(15) Interrupting pharmacological treatment, to what extent does it affect a person with ADHD in the following aspects?: emergency room visits, variations in the incidence of bicycle accidents, driving, work-related, etc.; alterations in relationships in the work and academic environment; legal complications.
(16) Upon initial diagnosis of ADHD, which treatment modality or modalities are shown to be most effective and safe in the long term?
(17) Among the following therapies: cognitive-behavioural, systemic, and psychodynamic, or combinations thereof, which one or what combination is most effective and safe for the treatment of ADHD?
(18) For a person with ADHD, what is the effectiveness and safety of training and education programmes aimed at

parents, family members, caregivers, and other people in the environment or of reference?
(19) What epidemiological data exist on ADHD, its evolution with the person's maturity, and potential risks that the individual will have to face?
(20) In the case of a girl or boy with symptoms of ADHD, what are the risks of not being diagnosed with ADHD?
(21) What should organisational skills training programmes include for boys and girls diagnosed with ADHD?
(22) What should behavioural training programmes for parents include as part of the first lines of treatment for boys and girls with ADHD in the preschool and school ages (4–11 years old)?
(23) What type of parental programmes should be included in the treatment of ADHD?
(24) How does ADHD evolve in individuals over the years?
(25) What are some of the symptoms of ADHD?
(26) How is ADHD diagnosed?

Appendix 2. Examples of questions and model answers

A.3. Clinical practice guideline on the management of depression in adults

(3) • **Question**: What are the perspectives of patients and relatives regarding depression and their experiences with the healthcare they received?
- **GPT-3's answer**: Qualitative studies have pointed out that patients and their families have a positive perception of the care received, although they have also expressed criticisms about the lack of time, lack of empathetic treatment, lack of coordination, and not taking the family into account. Additionally, patients emphasise the role of associations in helping patients with mental disorders, specifically those with depression.**Evaluation**:

(a) Coherence: 2.
(b) Veracity: 3.
(c) Evidence: 3.

- **LLaMA's answer**: In general, patients and their families have a limited understanding of depression and may feel fear, shame, or reluctance to seek help. They often feel frustrated with the healthcare system, which can be challenging to navigate and may provide inadequate or uneven care. Patients may experience a lack of empathy, time, and coordination in their medical care, which can lead to feelings of abandonment and discouragement. However, there are also positive testimonials from compassionate and effective healthcare professionals who have helped improve the quality of life for patients with depression.**Evaluation**:

(a) Coherence: 4.
(b) Veracity: 3.
(c) Evidence: 3.

- **LLaMA-2's answer**: The perspectives of patients and their families regarding depression and their experiences with the received healthcare are important for understanding how patients perceive their illness and how they value the care they receive. Qualitative studies have identified significant aspects of how patients perceive their illness, such as functional impact, the loss of pleasure in daily life, and difficulty in communicating their feelings to others. Moreover, experiences with the received healthcare can influence how patients access and interact with the healthcare system. Studies have identified that patients value empathy, understanding, and personalised care, as well as coordination among professionals and participation in decision-making. It has also been identified that family members may have specific support and guidance needs concerning the depression of their loved ones.**Evaluation**:

(a) Coherence: 5.
(b) Veracity: 3.
(c) Evidence: 4.

- **Valid answer**: In evaluating depression, it is recommended to consider the heterogeneity of its presentation, as well as the patient's perception of their symptoms and the disorder. The meaning and impact of depression on the patient's family should be explored, along with any potential needs that may arise, paying special attention to children, teenagers, and dependent family members under the care of the patient with depression. It is advisable to foster the communication of feelings and emotions in an empathetic and respectful environment. Additionally, the patient, and with their consent, their family and close friends, should play an active role in making decisions about treatment and developing a care plan.-Regarding their experiences with the health system, it is believed that the patient, due to the particularity of their illness, takes too long to seek medical attention. They have ambivalent feelings about the prescribed medications, being aware of their advantages and disadvantages. Concerning healthcare professionals, both patients and their families have common criticisms: lack of time, perceived lack of empathy, poor coordination, and not taking the family into account. Furthermore, they complain about the lack of close follow-up, insensitivity of professionals, limited time to attend to patients, and the absence of proper information about the disorder. Despite this, the involved professionals are generally well-regarded by patients.

(4) • **Question**: How should the evaluation for depression be conducted?
- **GPT-3's answer**: The evaluation of depression should be approached broadly and should not solely be based on symptom counting. Given that various factors can influence the development, course, and severity of depression, some areas to assess include the NICE guidelines, the clinical interview, scales and interviews with varying degrees of structure, and patients' perceptions of their illness and health services. Furthermore, the use of Whooley's two questions is recommended to identify cases of depression.**Evaluation**:

(a) Coherence: 4.
(b) Veracity: 3.
(c) Evidence: 4.

- **LLaMA's answer**: "The evaluation of depression should be approached broadly and should not solely be based on symptom counting. Some areas to assess include the sociodemographic, cultural, and biological factors that might affect the development or maintenance of depressive symptoms, as well as the patient's clinical history and response to treatment. The clinical interview is the essential procedure in diagnosing depression, and it is recommended to use scales and questionnaires to evaluate the severity and response to treatment.**Evaluation**:

(a) Coherence: 4.
(b) Veracity: 3.
(c) Evidence: 4.

- **LLaMA-2's answer**: The evaluation of depression should be approached broadly and not just based on symptom counting. Since various factors can influence the development, course, and severity of depression, some of the areas to assess include: Episode characteristics: duration, number, and intensity of symptoms, comorbidity. Psychosocial evaluation (social support and interpersonal relationships). Degree of associated dysfunction and/or disability. Past response to treatment. Safety assessment and risk of suicide.**Evaluation**:

(a) Coherence: 5.
(b) Veracity: 3.
(c) Evidence: 4.

- **Valid answer**: The assessment of depression should be conducted with a comprehensive approach and should not rely solely on symptom tallying. It is carried out through a clinical interview, which is the essential procedure, taking into account the information provided by the patient, relatives, and close acquaintances. Sometimes, to supplement the interview, questionnaires or tests are used, typically to learn more about the symptoms or the specific type of depression the individual might have. Given that there are various factors that can affect the onset, course, and severity of depression, some areas to evaluate include: characterising the episode, considering its duration, number, and intensity of symptoms, and any comorbidity; psychosocial assessment (social support and interpersonal relationships); the degree of dysfunction and/or associated disability; previous treatment response; and the risk of suicide.

(5) • **Question**: Which scales have the best psychometric properties for the evaluation of depression in adults?

- **GPT-3's answer**: The scales with the best psychometric properties for the evaluation of depression in adults are the Hamilton Rating Scale for Depression (HRSD or HAM-D), the Montgomery Asberg Depression Rating Scale (MADRS), and the BDI-II.**Evaluation**:

(a) Coherence: 5.
(b) Veracity: 3.
(c) Evidence: 5.

- **LLaMA's answer**: Among the most commonly used scales with good psychometric properties for the evaluation of depression in adults are the Hamilton Rating Scale for Depression (HRSD), the Montgomery Asberg Depression Rating Scale (MADRS), the Patient Health Questionnaire (PHQ-9), and the Beck Depression Inventory (BDI).**Evaluation**:

(a) Coherence: 5.
(b) Veracity: 3.
(c) Evidence: 5.

- **LLaMA-2's answer**: Some of the scales with the best psychometric properties for the assessment of depression in adults are the Hamilton Rating Scale for Depression (HRSD), the Montgomery Asberg Depression Rating Scale (MADRS), the Patient Health Questionnaire (PHQ-9), and the Beck Depression Inventory (BDI).**Evaluation**:

(a) Coherence: 5.
(b) Veracity: 3.
(c) Evidence: 5.

- **Valid answer**: Starting from the premise that the guide does not mention the best psychometric properties, but adequate psychometric properties, the most suitable scales are: The Beck Depression Inventory: both its original version and its updates BDI-IA and BDI-II; The Hamilton Rating Scale for Depression: there is an adaptation and validation in Spanish of its original version, and comparative analyses of the 6, 17, and 24 item versions; and the Montgomery Asberg Depression Rating Scale: there is a validated version in Spanish.

A.4. Clinical practice guideline on therapeutic interventions in attention deficit hyperactivity disorder

(3) • **Question**: Throughout the formative stages, in people with ADHD, what guidelines are appropriate to improve organisation and time management dedicated to school tasks and academic performance?

- **GPT-3's answer**: Expert groups agree in recommending that behavioural interventions should be a significant part of the treatment plan for ADHD at all ages, included in both programmes aimed at parental education and classroom training. These interventions must be adapted to each case based on the educational, behavioural, and emotional needs of a specific student, as well as the resources of each school. Moreover, it's crucial that parents, teachers, and individuals with ADHD receive proper education to understand the medical basis of the disorder and to see how a correct diagnosis would explain much of the behavioural difficulties and needs of these individuals. This education would help them view behavioural interventions as a step-by-step approach to developing skills that will enhance functionality in school, at home, or at work. That's why self-education in ADHD is recommended as a way to develop skills to adapt to the workplace.**Evaluation**:

(a) Coherence: 5.
(b) Veracity: 3.

(c) Evidence: 4.
- **LLaMA's answer**: The appropriate guidelines for improving organisation and time management dedicated to school tasks and academic performance in people with ADHD include:
(a) Routine and structured study plans.
(b) Organizational tools such as calendars, notebooks, and task lists.
(c) Auditory and visual stimuli to maintain attention on tasks.
(d) Short regular breaks for rest and recovery.
(e) Rewards for task completion and academic progress.
(f) Training in executive skills such as planning, working memory, and selective attention.
(g) Emotional and educational support to overcome challenges and obstacles.

Evaluation:

(a) Coherence: 4.
(b) Veracity: 3.
(c) Evidence: 4.
- **LLaMA-2's answer**: People with ADHD can benefit from specific interventions that help them improve their organisation, time management, and academic performance. These interventions may include planning techniques, time management strategies, self-control techniques, and social skills. Additionally, it's important for teachers and parents to work together to provide a safe and stimulating learning environment that allows the student with ADHD to develop their academic and social skills. As they grow older, it's important for students with ADHD to take more responsibility and autonomy in their own learning, which will help them develop greater self-esteem and self-confidence. Finally, periodic follow-ups are essential to ensure that students with ADHD continue progressing according to their individual goals.**Evaluation**:
(a) Coherence: 5.
(b) Veracity: 3.
(c) Evidence: 4.
- **Valid answer**: Individualized training, based on the degree of impact on the child, focuses on organisational skills and planning in school using new tools and routines for noting assignments and due dates. Organize schoolwork into folders, using checklists of the materials needed, and breaking tasks into steps, monitoring the time needed to complete them. Provide training for parents on information about the disorder, planning strategies, training for teachers on classroom management strategies, and training in new organisational skills for children. Application of cognitive-behavioural therapy specifically aimed at improving planning skills for students with ADHD. Ongoing monitoring and assessment of the management and planning over time is needed.

(4) - **Question**: During the educational stages of people with ADHD, and regardless of the type or combination of treatments, what kind of procedure or coordination model between the academic institution, the health system, and the family environment improves educational outcomes?
- **GPT-3's answer**: A collaborative intervention model, in which all stakeholders involved in the education process of individuals with ADHD participate, can increase their motivation to enhance their education. This would lead to an improvement in their deficits and academic outcomes. Furthermore, according to guidelines from the University of Michigan, parents, teachers, and individuals with ADHD need proper education to understand the medical foundation of the disorder. Recognizing an appropriate diagnosis would explain a large part of the behavioural difficulties and needs of these individuals. This education would help them view behavioural interventions as a step-by-step approach to skill development that will enhance their functionality at school, home, or work.
- **LLaMA's answer**: The studies found suggest that an effective coordination model among the academic institution, the healthcare system, and the family environment is crucial for enhancing educational outcomes of individuals with ADHD. This coordination model may encompass effective communication among teachers, parents, physicians, and other professionals involved in the individual's health and learning. The research recommends the establishment of a link coordinated by the school doctor to foster collaboration around the children's challenges at school, the progression of their symptoms, and updating pedagogical strategies when needed. This coordination model can also involve the creation of individualised educational support plans tailored to the specific needs of the individual with ADHD.**Evaluation**:
(a) Coherence: 5.
(b) Veracity: 3.
(c) Evidence: 5.
- **LLaMA-2's answer**: While I don't have specific information on the coordination procedure or model that enhances educational outcomes in individuals with ADHD, it has been identified that collaboration among various stakeholders such as families, teachers, and healthcare professionals can improve treatment quality and academic results. Studies have shown that a collaborative intervention model, involving all stakeholders engaged in the educational process, can increase the motivation of individuals with ADHD to enhance their training, which would result in an improvement in their deficits and academic outcomes.**Evaluation**:
(a) Coherence: 4.
(b) Veracity: 3.
(c) Evidence: 3.
- **Valid answer**: Individualized training, depending on the degree of impairment of the child, in organisational skills, and planning in school in new tools and routines to note tasks and delivery dates. Organize school assignments in folders, using checklists of the materials needed, as well as breaking tasks down into steps, monitoring the time needed to complete them. Provide training for parents in information about the disorder, planning strategies, training for teachers in classroom management strategies, and training in new organisational skills for children. Application of cognitive-behavioural therapy specifically focussed

on improving the planning skills of students with ADHD. It is necessary to track the management and planning carried out over time.

(5) • **Question**: Are interventions for training and developing social skills effective and safe for people with ADHD, regardless of any associated comorbidities they may have, to integrate these individuals into the labor market or, once integrated, to allow them to maintain appropriate relationships in the workplace?
- **GPT-3's answer**: I do not have the information to answer that question.**Evaluation**:
  (a) Coherence: 1.
  (b) Veracity: 1.
  (c) Evidence: 1.
- **Hermes (LLaMA)'s answer**: I do not have the information to answer this question based on the context.**Evaluation**:
  (a) Coherence: 1.
  (b) Veracity: 1.
  (c) Evidence: 1.
- **LLaMA-2's answer**: I do not have the information to answer that question.**Evaluation**:
  (a) Coherence: 1.
  (b) Veracity: 1.
  (c) Evidence: 1.
- **Valid answer**: In general, they are safe and effective, although this depends on factors such as the level of disease impact and the life stage at which the patient begins receiving therapies. There is a direct relationship between the effectiveness of the treatment and when it begins. Therefore, the sooner treatments are started, the better the results obtained. Thus, individuals with ADHD need appropriate education to understand the medical basis of the disorder, which would help them see behavioural interventions as a step-by-step approach to skill development that will assist them in improving their job performance. Self-education on ADHD is recommended as a means of developing skills to adapt to the workplace. The HAS Guide recommends interventions over not having them in all studied aspects, one of which refers to employment. It is highlighted that strategies learned through cognitive-behavioural therapy (CBT) can help people with ADHD establish a positive and realistic mindset, allowing them to overcome or minimise long-held negative messages and manage setbacks and problems that would otherwise be viewed as signs of failure or incompetence. Likewise, through self-education, they can develop strategies that consider the strengths and weaknesses of individuals with ADHD when choosing a profession, maintaining their social relationships, and being able to keep their jobs. Many of the adaptations that allow people with ADHD to develop their potential outside the workplace can also be applied in work situations, such as organisation and planning strategies in daily life.