Physio: An LLM-Based Physiotherapy Advisor

Rúben Almeida[⋆1][0000−0002−1942−2399], Hugo Sousa[⋆1,2][0000−0003−3226−9189], Luís F. Cunha[⋆1,2][0000−0003−1365−0080], Nuno Guimarães[1,2][0000−0003−2854−2891], Ricardo Campos[1,3,4][0000−0002−8767−8126], and Alípio Jorge[1,2][0000−0002−5475−1382]

[1] INESC TEC, Porto, Portugal
[2] University of Porto, Porto, Portugal
[3] University of Beira Interior, Covilhã, Portugal
[4] Ci2 - Smart Cities Research Centre, Tomar, Portugal
{ruben.f.almeida,luis.f.cunha,hugo.o.sousa,
nuno.r.guimaraes,ricardo.campos,alipio.jorge}@inesctec.pt

Abstract. The capabilities of the most recent language models have increased the interest in integrating them into real-world applications. However, the fact that these models generate plausible, yet incorrect text poses a constraint when considering their use in several domains. Healthcare is a prime example of a domain where text-generative trustworthiness is a hard requirement to safeguard patient well-being. In this paper, we present Physio, a chat-based application for physical rehabilitation. Physio is capable of making an initial diagnosis while citing reliable health sources to support the information provided. Furthermore, drawing upon external knowledge databases, Physio can recommend rehabilitation exercises and over-the-counter medication for symptom relief. By combining these features, Physio can leverage the power of generative models for language processing while also conditioning its response on dependable and verifiable sources. A live demo of Physio is available at https://physio.inesctec.pt.

Keywords: Retrieval-augmented generation · Information extraction · Conversational health agents

1 Introduction

Although language models (LMs) have long been studied by the research community, they only reached mainstream attention with the release of the ChatGPT application by OpenAI [3]. This application granted the public access to a highly effective generative model, GPT-3.5 [15], that was capable of producing coherent conversations on various topics, a novelty at the time. This development naturally led to the emergence of numerous applications and discussions regarding the potential applications of generative models in various domains, such as law [16], education [17], and health [11,12]. However, these models also exhibited significant limitations that hindered their implementation in those domains. At the

⋆ Equal contribution.

top of that list is the hallucination problem [14], *i.e.*, their propensity to generate incorrect yet convincing answers. This limitation prompted increased research into grounding the text generated by these models on reliable sources, a task known as retrieval-augmented generation [13]. The general approach starts by retrieving documents relevant to the input query and subsequently using them to generate an answer. By doing so, one can link the generated texts to the original documents, thereby providing the user references where he/she can get more information supporting the generated answer [9,10]. This research gave rise to systems like BingChat [2] and Bard [1], search engines that combine the personalization of answers generated by LMs with the trustworthiness provided by the retrieval component of the system. This concept can be taken one step further to be applied to domain-specific applications by constraining the retrieval component to a specialized set of documents. This is the main idea behind the demo we present in this paper, Physio, a chat-based application tailored to help users in physical rehabilitation. Physio's answers are generated from an augmented prompt that compiles the user input with documents from a curated knowledge base specifically tailored to only contain reliable sources for physical rehabilitation. Furthermore, the sentences in the generated answer present references to those documents. Apart from that, Physio's answer can also contain exercise and over-the-counter medication recommendations whenever appropriate. The source code for Physio is open-source and available on GitHub[5].

2 Physio

To some extent, Physio can be thought of as an initial step toward an artificial intelligent physiatrist. To this end, we designed it to offer functionalities akin to those found in a physiatrist appointment, specifically: explaining/discussing the problem that the user is querying; and recommending exercises and medication when deemed appropriate. For the LM, Physio relies on OpenAI GPT-4 model which is the most effective LM at the time of development of this demo [14]. An example answer from Physio is illustrated in Figure 1 for the prompt "I feel pain in my lower back. What can I do?" In the following sections, we explain how Physio's answer is built.

2.1 Knowledge-base Construction

The construction of the knowledge for Physio commenced by scraping the Rehab Hero website [6]. This website boasts an extensive compilation of exercises, each accompanied by an instructional video demonstrating execution, and information on the specific physical conditions they address. After scraping this website, we compiled all the physical conditions that we found on the Rehab Hero pages and queried each of them in a set of 20 websites that provided more information about the condition[6]. To ensure the reliability of our system the list of websites

[5] https://github.com/liaad/physio

[6] More details about the list of websites selected can be found on the GitHub repository.

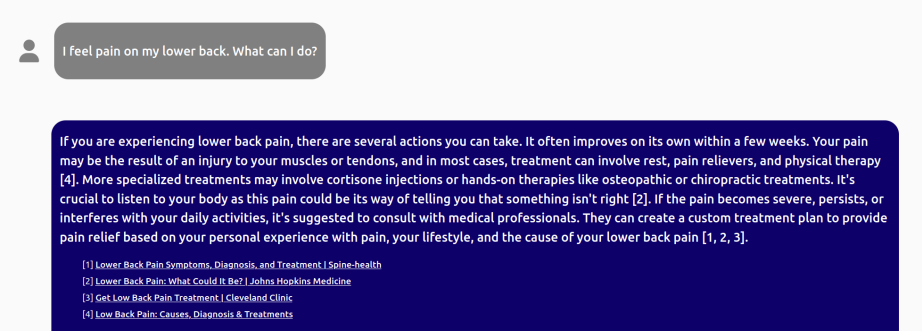

Fig. 1: Screenshot from Physio web demonstration. The user input is in the grey box, while the system answer is presented in the blue box.

was curated and validated by a physiotherapist and includes sources like Mayo Clinic [4], NHS [7], and OrthoInfo [5]. For the medication-related aspects, we rely on the DrugBank database [18]. The information obtained was then indexed in a MongoDB database composed of three collections: `exercises`, `webpages`, and `medications`. The first two are indexed based on the associated condition, while the last is indexed by drug name.

2.2 Data Pipeline

When a user submits a query in Physio, the text undergoes processing through a data pipeline to generate a response. The initial step in this pipeline is to **verify if it is an English physiotherapy-related prompt**. This validation is achieved by using the LM and a predefined validation prompt template that assesses whether the user's input is related to physiotherapy and written in English. The validation prompt template conditions the LM's response to produce a boolean output (either "True" or "False") so that it can be programmatically interpreted. If the input is deemed invalid, the system provides a default response. Instead, if the input is validated, the application proceeds to the **condition identification** step, where it determines the condition in the user's question. This is accomplished by employing a few-shot template to instruct the LM to identify the condition. For instance, when presented with the input "I have sprained my ankle" the model should identify the condition as "ankle sprain."

Once the condition is identified, it is **linked to one of the entries in our database**. This linkage process first attempts an exact match, followed by a search in the list of aliases (for instance "lumbago" is in the list of aliases for "back pain"), and, as a last resort, employs substring matching. In case no match is found, the LM is prompted with the user query, and the answer is returned. Otherwise, the pipeline advances to the generation of the answer and extraction exercises and medication.

Answer Generation With the linked condition one can retrieve the list of documents related to that condition from the `webpages` collection. Among the pages available, we employ the BM25 retrieval model [8] to search and rank them based on their relevance to the user's input. The top five ranked documents are subsequently provided to the generative model, along with the user's input query, using a prompt template designed to instruct the model to answer user questions using the information contained in these pages.

To provide the user a way to verify the trustworthiness of the generated text, a list of references is incorporated by ranking the sentences from the original source pages in relation to the sentences in the generated text, again, using the BM25 ranking method. Note that determining the optimal number of references to include in the final answer is not trivial. In our application, we establish a heuristic to use as reference the top-N ranked sentence-document pairs, where N is the number of generated sentences.[7] However, the final answer may not necessarily contain the same number of references as sentences, as a generated sentence can be highly similar to multiple sentences within a given document.

Exercise & Medication Extraction The linked condition is also used to fetch the exercises directly from the `exercises` collection, as they are indexed by condition. For the linked condition we randomly sample up to five exercises to be presented in the web interface.

The final element of the response pertains to medication recommendations. This is accomplished by instructing the LM to provide medication suggestions based on the user's query, the linked condition, and the generated answer. The prompt explicitly specifies that the response should be in the form of a JSON-parsable list of strings, where each string represents a medication. After parsing these strings, we conduct a search for the recommended drugs within the `medication` collection of our database, first by exact matching and subsequently with fuzzy matching.

The last task of the pipeline is to combine the three components of the answer, send them to the frontend, and cache the result in the database.

Ethical Considerations Given the sensitive nature of this domain, ethical considerations are paramount. As a result, we include a disclaimer on Physio's website, explicitly stating that it is a research demonstration, and we strongly advise users to consult with a specialist before making any decisions regarding their health. Furthermore, we have limited medication recommendations to include only over-the-counter options.

Aknowledgments This work is financed by National Funds through the Fundação para a Ciência e a Tecnologia, within the project StorySense (DOI 10.54499/2022.09312.PTDC) and the Recovery and Resilience Plan within the scope of the Health From Portugal project.

[7] While this heuristic has been effective in practice, ongoing research is aimed at refining the reference selection process based on similarity scores.

References

1. Bard. https://bard.google.com/chat, accessed: 2023-10-05
2. Bing chat. https://www.microsoft.com/en-us/edge/features/bing-chat, accessed: 2023-10-05
3. Chatgpt. https://chat.openai.com/, accessed: 2023-10-05
4. Mayo clinic. https://www.mayoclinic.org/, accessed: 2023-10-05
5. Orthoinfo. https://orthoinfo.aaos.org/, accessed: 2023-10-05
6. Rehab hero. https://www.rehabhero.ca/, accessed: 2023-10-05
7. United kingdom national health service. https://www.nhs.uk/, accessed: 2023-10-05
8. Amati, G.: BM25, pp. 257–260. Springer US, Boston, MA (2009), https://doi.org/10.1007/978-0-387-39940-9_921
9. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling large language models to generate text with citations (2023)
10. Huang, J., Chang, K.C.C.: Citation: A key to building responsible and accountable large language models (2023)
11. Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., Tseng, V.: Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health **2**(2), e0000198 (Feb 2023). https://doi.org/10.1371/journal.pdig.0000198
12. Levine, D.M., Tuwani, R., Kompa, B., Varma, A., Finlayson, S.G., Mehrotra, A., Beam, A.: The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model . https://doi.org/10.1101/2023.01.30.23285067
13. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
14. OpenAI: Gpt-4 technical report (2023)
15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022)
16. Savelka, J.: Unlocking Practical Applications in Legal Domain: Evaluation of GPT for Zero-Shot Semantic Annotation of Legal Texts (May 2023). https://doi.org/10.1145/3594536.3595161, arXiv: 2305.04417
17. Savelka, J., Agarwal, A., Bogart, C., Song, Y., Sakr, M.: Can Generative Pre-trained Transformers (GPT) Pass Assessments in Higher Education Programming Courses? (Mar 2023), http://arxiv.org/abs/2303.09325, arXiv: 2303.09325
18. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., Wilson, M.: DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research **46**(D1), D1074–D1082 (Jan 2018). https://doi.org/10.1093/nar/gkx1037