Towards Automatically Filling Questionnaires from Clinical Records with Large Language Models

Valeria Nardoni[1], Marco Lippi[1,*], Giulia Hyeraci[2], Martina Maccari[3], Amirreza Dehghan Tarazjani[6], Gianni Virgili[4,5], Rosa Gini[2] and Simone Marinai[1]

[1]*Department of Information Engineering, University of Florence, via di Santa Marta 3, 50139 Florence, Italy*
[2]*ARS Toscana, via Dazzi 1, 50141 Florence, Italy*
[3]*SOD Ottica Fisiopatologica – Clinical Trial Center, Florence, Italy*
[4]*Department of Neurosciences, Psychology, Drug Research and Child Health (NEUROFARBA), University of Firenze, Florence, Italy*
[5]*Ophthalmology, IRCCS-Fondazione Bietti, Rome, Italy*
[6]*University Medical Center Utrecht*

Abstract

The use of large language models in healthcare is rapidly growing. In this work, we are interested in analyzing the capabilities of large language models in filling questionnaires related to clinical practices, where the information needed to answer each specific question is contained in the clinical records of a given patient. We present preliminary experiments on a publicly available dataset in the English language, with very promising results that show the great potential of the approach and motivate further research in this challenging direction.

Keywords

Large language models, clinical records, questionnaire filling, prompt engineering

1. Introduction

In the last couple of years, Large Language Models (LLMs) have produced a deep revolution in the field of artificial intelligence, and their impact is notoriously spreading all over a wide variety of application domains. One major area of interest is clearly healthcare, where LLMs are already contributing to the development of services for clinicians, companies and public administrations [2].

In this work, we are interested in a specific task in healthcare: that of exploiting LLMs to automatically fill questionnaires related to the clinical records of patients. Clinical questionnaires are a precious source of information for researchers and clinicians, as they can be used to identify patterns and relations across the characteristics of a given patient and his/her clinical history. As LLMs are nowadays one of the leading solutions for natural language processing (NLP) problems, we propose an approach that naturally builds on this new technology in order to answer the questions via prompting.

We derived a data extraction questionnaire from the SeValid study protocol, which has been developed by the international VAccine monitoring Collaboration for Europe non-profit association (VAC4EU) in collaboration with AOU Careggi (the largest hospital in the area of Florence, Italy), and the University Medical Center in Utrecht. The questionnaire used in this study aims to confirm the diagnosis of several conditions, among which deep vein thrombosis (DVT). The protocol will also adopt human verification after electronic medical records coding, in order to map the occurrence of these conditions. The developed questionnaires are aimed at identifying whether admission or hospitalization was motivated by the occurrence of DVT. The development of this methodology could trigger retrospective studies for the analysis of links between risk factors and certain pathologies.

NL4AI 2024: Eighth Workshop on Natural Language for Artificial Intelligence, November 26-27th, 2024, Bolzano, Italy [1]

[*] Corresponding author.

valeria.nardoni@edu.unifi.it (V. Nardoni); marco.lippi@unifi.it (M. Lippi); giulia.hyeraci@ars.toscana.it (G. Hyeraci); maccarim@aou-careggi.toscana.it (M. Maccari); dehghantarazjani@vac4eu.org (A. D. Tarazjani); gianni.virgili@unifi.it (G. Virgili); rosa.gini@ars.toscana.it (R. Gini); simone.marinai@unifi.it (S. Marinai)
0000-0002-9663-1071 (M. Lippi); 0000-0002-6536-2083 (G. Hyeraci); 0009-0002-2593-4267 (M. Maccari); 0000-0002-1238-4107 (A. D. Tarazjani); 0000-0002-9960-2989 (G. Virgili); 0000-0002-6250-877X (R. Gini); 0000-0002-6702-2277 (S. Marinai)

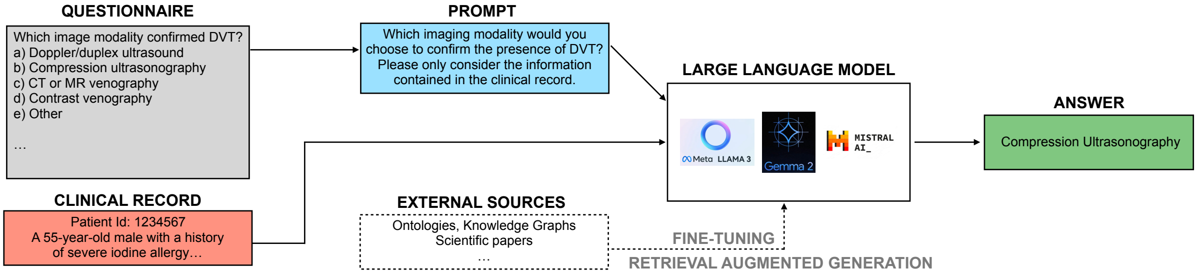
QUESTIONNAIRE
Which image modality confirmed DVT?
a) Doppler/duplex ultrasound
b) Compression ultrasonography
c) CT or MR venography
d) Contrast venography
e) Other
...
PROMPT
Which imaging modality would you choose to confirm the presence of DVT? Please only consider the information contained in the clinical record.
LARGE LANGUAGE MODEL
Meta LLAMA 3
Gemma 2
MISTRAL AI_
ANSWER
Compression Ultrasonography
CLINICAL RECORD
Patient Id: 1234567
A 55-year-old male with a history of severe iodine allergy…
EXTERNAL SOURCES
Ontologies, Knowledge Graphs
Scientific papers
...
FINE-TUNING
RETRIEVAL AUGMENTED GENERATION

Figure 1: Main architecture and workflow of our system. The questions contained in the questionnaire (gray) are transformed into a prompt (blue), that is fed as input to the LLM together with a clinical record (red) to obtain an answer (green) to a specific clinical question. The bottom part (in white and dashed) is currently not implemented in our system, but will be used in our future work.

2. Methodology

The adopted methodology is illustrated in Figure 1. We analyzed a set of questionnaires, provided by the Regional Agency for Healthcare in Tuscany (ARS Toscana) and the major hospital in the area of Florence (AOU Careggi), whose filling requires the retrieval of relevant information from clinical records. The selected questions are manually transformed into specific prompts that are subsequently fed to an LLM together with the clinical record.

Different strategies can be designed according to different levels of interaction with the LLM. As a first strategy, the prompt could be simply formatted in a zero-shot setting (i.e., without any example or other information beside the question) or in a few-shot setting (i.e., by adding some cases and examples to the prompt). An additional modification to the prompt would require to ask the model to also motivate its answer, for example by reporting evidence from the clinical record. In fact, it is also a well-known fact that the way in which the prompt is phrased can have a strong impact on the performance of the LLM [3]. Just to make an example, multiple choice questions or closed-end questions can be phrased as collections of questions that require just a yes/no answer. Similarly, different prompts can ask the LLM to produce in output just the plain answer, or even the reasoning process, that typically includes the full Chain-of-Thought [4].

A more advanced strategy could also be explored, by adding the possibility to fine-tune the LLM with data coming from a collection of clinical records, or even to perform retrieved-augmented generation [5] using external knowledge bases such as ontologies, knowledge graphs, or scientific papers (see bottom part of Figure 1). Nevertheless, these solutions require more resources, and we plan to investigate them in future research. Our preliminary experimental evaluation, described in Section 4, will compare three different open-source LLMs (namely, *Gemma2-9B*, *LLaMa3-8B*, and *Mistral-7B*) with various prompting strategies. Adopting larger models in future experiments will likely improve performance, as it is customary with LLMs.

3. Dataset

To test our approach, we used the publicly available PMC-patients dataset [6][1]. The PMC-Patients dataset represents a significant resource for clinical research and data analysis, encompassing anonymized patient summaries, demographic data, and relational annotations. It is divided into three main sections: (i) the actual PMC-Patients dataset, including the primary clinical data and detailing clinical cases, (ii) the ReCDS benchmark, providing metrics for evaluating Retrieval-based Clinical Decision Support (ReCDS) systems, (iii) meta data, containing JSON-formatted metadata that offer additional information on clinical notes, relevant scientific articles, and similar patients.

The dataset comprises 167,000 clinical summaries, associated with 3.1 million relevant patient-article pairs and 293,000 similar patient-patient pairs. It serves both as a collection of high-quality and diverse

[1] https://pmc-patients.github.io/

clinical cases on a large scale and an extensive resource for benchmarking ReCDS systems. Each case is linked to real patients described in articles from PubMed Central (PMC), a free repository of scientific articles in biomedicine and life sciences. The information in the dataset consists of fully anonymized clinical notes and relational annotations, collected through an automated process that uses regular expressions to identify specific patterns in the text of articles published on PMC. This process extracts relevant sections describing patients in detail, such as clinical case reports. After identifying the relevant sections, clinical summaries are extracted along with demographic data, including age and gender. The summaries are then filtered to exclude those that are too brief, not in English, or lacking demographic information. Finally, the selected summaries are linked to related articles and similar patients through citation relationships between articles within PubMed.

In order to select interesting and challenging case studies, we processed and analyzed the patient data with advanced NLP libraries such as NLTK for text tokenization, stopword removal, and lemmatization. A text preprocessing function was defined to perform several operations, including: converting the text to lowercase for consistency, removing punctuation, tokenizing the text, breaking it into individual words, removing stopwords to reduce noise in the text, and lemmatizing words to reduce them to their base form. This preprocessing function was applied to the initial dataset to select two different scenarios. In a first scenario, we selected a tiny subset of patients for initial investigations: we collected 41 patients related to the keyword "Hemoperitoneum", further reduced to 35 by removing cases that involved more patients in a single clinical study. This dataset was used in the first case study illustrated in Section 4.1. In a second scenario, a list of specific keywords related to deep vein thrombosis (DVT) was created and used to identify patients whose records mentioned DVT. Additionally, starting from the PMC-Patients dataset, text mining methods were used to extract all cases mentioning DVT. This process led to the selection of 1,726 patients affected by DVT. The second and third case studies described in Sections 4.2.1 and 4.2.2 exploit this dataset.

4. Experiments

We considered three different case studies that cover questions that require different types of answer: (1) reporting the blood pressure measurement mentioned in the clinical record, if any; (2) identifying in the clinical record the imaging technique used to diagnose DVT, if any; (3) providing the localization in the patient's body of the DVT (upper, lower or other). In the first case, a numerical value has to be provided; in the second case, one of five possible alternatives has to be selected; finally, in the third case, a task with three categories is performed. We used three different LLMs via the Ollama library:[2] Gemma2-9B, Llama3-8B, and Mistral-7B. We did not perform any change in the default parameters of the models, while we concentrated our analysis on the way in which the prompts could be formulated.

4.1. Blood pressure measurements

The first case study was chosen for its simplicity, and was conducted on the limited dataset of 35 patients described in Section 3. Here, we simply asked the LLM to retrieve from the clinical record the blood pressure values in mmHg, in the format systolic/diastolic. This is very common information, which serves as a crucial reference point in many clinical decisions.

We compared the three chosen LLMs in the zero-shot learning and few-shot learning settings. In some prompts, we also asked the LLM to explain its answer. Even in this simple scenario, we observed how the formulation of the prompt has a very strong impact on performance. In Table 1 we report the prompt that achieves the best performance, i.e., all the replies are correct, for all the three LLMs. This setting corresponds to few-shot learning, without asking for further explanations from the model.

[2] https://ollama.com/

Table 1
Best prompt used for the experiment on blood pressure measurement (few-shot learning, no chain-of-thought).

Prompt:
Your task is to extract the blood pressure readings (systolic/diastolic in mmHg) from the given medical records. If the blood pressure readings are not present, respond with 'Null'. Base your response solely on the text provided in the patient record. Here are some examples:
– Patient record: The patient has a history of hypertension but no current measurements available. Response: Null
– Patient record: The patient was seen for a regular check-up. Blood pressure was 116/89 mmHg. Response: 116/89
– Patient record: No significant changes in patient's vitals. Blood pressure stable. Response: Null
Now, here is the new patient record: {record}
Based on the text, respond only with the value or with Null if not available.

4.2. Deep vein thrombosis

The second and third case studies are related to the much more challenging scenario of studying DVT, a pathology that involves blood clotting in deep veins, rather than superficial ones. This phenomenon is a complex condition with many factors involved, including both acquired and inherited predispositions, and environmental factors. We consider two different questions that are important from a clinical perspective: (i) to identify the imaging technique used to diagnose DVT, and (ii) to perform a rough localization (upper vs. lower) of the DVT in the patient's body.

4.2.1. Imaging technique identification

Different imaging techniques can be used to diagnose DVT in patients: namely, compression ultrasonography, CT or MR venography, contrast venography, Doppler/Duplex ultrasound, or others. This information is very important, as it can be related to the clinical condition of the patient: for example, the use of contrast agents, should be avoided in patients with renal failure, a high risk of contrast-induced nephropathy, or allergies. Also for this scenario, we compared the zero-shot and few-shot settings, even with the additional requirement of providing evidence to justify and motivate the generated answers. Table 3 shows an example of the prompt we adopted for the few-shot learning setting, when asking for explanations and evidence in the answer.

4.2.2. Localization

The final scenario that we consider deals with the localization of the DVT, which can be either detected in the upper part or in the lower part of the body. The third alternative is that no information is provided within the clinical record. Again, we consider the same settings as in the previous scenarios.

4.3. Discussion

The experimental results consistently indicate that, among the analyzed models, Gemma2 stands out as the best performing, most likely due to larger number of parameters (9.24 billion). In addition, we observed that requiring the model to provide evidence to support responses systematically improves the results, confirming the importance of an approach that stimulates transparent reasoning, as it happens for Chain-of-Thought prompts [4]. Llama3, with its 8.03 billion parameters, showed a fair ability to handle complex tasks, outperforming the Mistral model (7.25 billion parameters) in some cases. However, despite the larger number of parameters, Llama3 did not always provide superior performance, highlighting that model size is not the only determining factor for success. We also notice that few-shot learning does not always improve performance: this behaviour suggests that additional work on the prompting strategy could produce better results.

Furthermore, we observe that the results of the third experiment, that on DVT localization, were significantly better, as the classification was limited to fewer categories, and thus less prone to misinterpretation. For example, in questions regarding the location of DVT (upper or lower body part),

Table 2

Accuracy for different LLMs and settings (zero-show vs. few-shot), possibly with the use of evidence requests during prompt, for the tasks of DVT imaging technique identification (ITI) and localization (LOC).

LLM	Setting	Evidence	Accuracy (ITI)	Accuracy (LOC)
Gemma2	Zero-shot	No	72.3	75.9
	Zero-shot	Yes	78.3	90.1
	Few-shot	No	67.7	80.5
	Few-shot	Yes	74.6	85.2
Llama3	Zero-shot	No	63.9	69.2
	Zero-shot	Yes	66.1	88.8
	Few-shot	No	60.9	75.8
	Few-shot	Yes	75.2	76.6
Mistral	Zero-shot	No	70.6	66.3
	Zero-shot	Yes	70.9	78.4
	Few-shot	No	60.1	51.3
	Few-shot	Yes	75.4	52.2

Table 3

Example of prompt for the few-shot learning setting, asking for evidence in the answer.

Prompt:
When selecting an imaging modality to confirm deep vein thrombosis (DVT), it is important to consider the patient's record. Below are some examples of how different imaging techniques were chosen based on the clinical scenarios.
Example 1:
Patient Record: A 70-year-old female with a history of renal failure presents with pain and swelling in her right leg. Due to her high risk for contrast-induced nephropathy, Doppler/Duplex Ultrasound was used to confirm the presence of DVT.
Response: Doppler/Duplex Ultrasound
Explanation: The choice of Doppler/Duplex Ultrasound avoids the use of contrast agents, which is crucial given the patient's renal failure and high risk of contrast-induced nephropathy.
Citation: 'Doppler/Duplex Ultrasound was used to confirm the presence of DVT.'

Example 2:
Patient Record: A 55-year-old male with a history of severe iodine allergy presents with pain and redness in his left leg. Given the contraindication to iodine-based contrast agents, Compression Ultrasonography was selected to confirm DVT.
Response: Compression Ultrasonography
Explanation: Compression Ultrasonography is a reliable method that does not require the use of contrast agents, making it suitable for a patient with a severe iodine allergy.
Citation: 'Compression Ultrasonography was selected to confirm DVT.'

Example 3:
Patient Record: A 60-year-old male with no known allergies or renal issues, but with a history of recurrent DVT, presents with new swelling in his left calf. CT Venography was chosen to provide detailed imaging.
Response: CT or MR venography.
Explanation: CT Venography provides a high-resolution image that is useful in complex cases like recurrent DVT, where detailed visualization of the venous system is necessary.
Citation: 'CT Venography was chosen to provide detailed imaging.'

Now, please analyze the following patient record: patient_record Which imaging modality would you choose to confirm the presence of DVT? Please start with your choice (e.g., 'Response: Doppler/Duplex Ultrasound'), then extract and quote the specific details from the patient record that led to your decision (e.g., 'Citation: [text]'), and finally explain your reasoning.
Options: compression ultrasonography; CT or MR venography; contrast venography; Doppler/Duplex ultrasound; Other

the clarity of the task facilitated correct identification by LLMs, as there were no subtle nuances or semantic ambiguities present in the clinical records. Conversely, in more complex scenarios, as in the second task of determining the diagnostic technique used to confirm DVT, the models showed greater difficulty. Overlap between similar techniques, such as Doppler Ultrasound and Compressive Ultrasonography, often emerged, making it complex to distinguish between the two methodologies in poorly detailed texts. A common error in ITI occurs when the LLM provides an answer with an imaging technique mentioned in the clinical report, even though it is not the one actually used for diagnosis. As an example, the following excerpt from a clinical report describes both Doppler ultrasound and compressive ultrasonography, but only one of these was actually used to diagnose deep vein thrombosis (DVT): the model thus incorrectly selects the unused technique, confusing the clinical context.

> *The patient presented with acute dyspnea and chest tightness suggestive of pulmonary embolism (PE). While a CTPA ultimately confirmed the diagnosis of PE, Doppler/Duplex Ultrasound is the more appropriate initial study for suspected DVT. This is because it is non-invasive, readily available, and cost-effective. The record mentions elevated D-dimer levels and signs of right ventricular strain on POCCUS (portable echocardiography). POCCUS helps assess right ventricular function, but Doppler/Duplex Ultrasound is the gold standard for visualizing deep veins and detecting blood clots (DVT), which can lead to PE.*

In other cases, the LLM attempts to infer which methodology would be most appropriate for the patient based on general criteria, rather than sticking strictly to the information provided in the report. This behaviour leads to answers that, although plausible, do not reflect the reality documented in the clinical report.

5. Conclusions

Filling questionnaires regarding the clinical history of patients is an important practice that can provide crucial information to identify links between risk factors and pathologies. Extracting the relevant information from the clinical records requires significant effort and time from domain experts, which makes it very hard to conduct this kind of study on a very large scale. In this paper, we argue that LLMs could be exploited for this task, due to their capabilities of natural language comprehension.

Acknowledgments

We thank the SeValid project (Validation on the identification of myocarditis and deep vein thrombosis cases from the hospital admission records of Careggi Hospital) promoted by ARS Toscana and Neurofarba, co-promoted by the “Vaccine monitoring collaboration for Europe” organization (VAC4EU), and coordinated by the Physiopathological Optics Department, AOU Careggi. This research has been partially funded by CAI4DSA actions (Collaborative Explainable neuro-symbolic AI for Decision Support Assistant), of the FAIR national project on artificial intelligence, PE 1 PNRR (https://fondazione-fair.it/).

References

[1] G. Bonetta, C. D. Hromei, L. Siciliani, M. A. Stranisci, Preface to the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Eighth Workshop on Natural Language for Artificial Intelligence (NL4AI 2024) co-located with 23th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2024), 2024.

[2] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large language models in medicine, Nature medicine 29 (2023) 1930–1940.

[3] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 (2023).

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, Advances in NeurIPS 35 (2022) 24824–24837.

[5] P. Lewis, E. Perez, A. Piktus, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[6] Z. Zhao, Q. Jin, F. Chen, T. Peng, S. Yu, A large-scale dataset of patient summaries for retrieval-based clinical decision support systems, Scientific data 10 (2023) 909.