dHealth 2024
D. Hayn et al. (Eds.)
© 2024 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI240006

Large Language Model-Based Evaluation of Medical Question Answering Systems: Algorithm Development and Case Study

Daniel REICHENPFADER[a,1], Philipp RÖSSLHUEMER[b] and Kerstin DENECKE[a]
[a] *Bern University of Applied Sciences, Biel/Bienne, Switzerland*
[b] *Department of Diagnostic, Interventional and Pediatric Radiology, Bern University Hospital, University of Bern, Bern, Switzerland*

Abstract. Background: Healthcare systems are increasingly resource constrained, leaving less time for important patient-provider interactions. Conversational agents (CAs) could be used to support the provision of information and to answer patients' questions. However, information must be accessible to a variety of patient populations, which requires understanding questions expressed at different language levels. Methods: This study describes the use of Large Language Models (LLMs) to evaluate predefined medical content in CAs across patient populations. These simulated populations are characterized by a range of health literacy. The evaluation framework includes both fully automated and semi-automated procedures to assess the performance of a CA. Results: A case study in the domain of mammography shows that LLMs can simulate questions from different patient populations. However, the accuracy of the answers provided varies depending on the level of health literacy. Conclusions: Our scalable evaluation framework enables the simulation of patient populations with different health literacy levels and helps to evaluate domain specific CAs, thus promoting their integration into clinical practice. Future research aims to extend the framework to CAs without predefined content and to apply LLMs to adapt medical information to the specific (health) literacy level of the user.

Keywords. Natural Language Processing, Consumer Health Information, Algorithms, Conversational Agents, Large Language Model

1. Introduction

In healthcare systems facing escalating resource constraints, conversational agents (CAs) represent a significant opportunity to improve patient-provider interactions and streamline information exchange. CAs are "computer programs designed to engage in human-like conversations with users" [1]. They can be grouped according to the direction of information flow: Information can flow either from the patient to the provider, e.g. facilitating history taking, symptom checking and triage, or from the provider to the patient, e.g. facilitating response to patient queries. In addition, CAs can support bi-directional information exchange. CAs designed to answer user questions are related to question-answering (QA) systems. These can be further categorized based on their

[1] Corresponding Author: Daniel Reichenpfader, Institute for Patient-centered Digital Health, Bern University of Applied Sciences, Biel/Bienne, Switzerland, E-Mail: daniel.reichenpfader@bfh.ch

technical implementation: On the one hand, patient questions can be mapped to the most similar example of pre-defined Question-Answer-Pairs (QAPs), either based on intent recognition or similarity-based techniques. With pre-defined QAPs, clinicians have absolute control over the provided content and answer generation is transparent, reducing the risk of providing wrong answers. On the other hand, Large Language Models (LLMs) might be used, including retrieval-augmented generation (RAG) [2] or immediate generation of an answer by the model itself [3]. Recent research shows that LLMs might provide comparably accurate and less biased answers to patients' medical questions compared to human experts, although still having limitations, e.g. interpretability of answers [3]. Therefore, many existing health CAs are using the approach with predefined QAPs.

In a setting, where a CA is supposed to answer patient queries, it is essential that it can handle a diverse set of user input. In healthcare, this means that there are users with a high health literacy and knowledge of medical terms, and those with low health literacy. Health literacy is defined as "the ability of an individual to obtain and translate knowledge and information in order to maintain and improve health in a way that is appropriate to the individual and system contexts" [4]. Pre-defined QAPs can never cover all possible wordings of user inputs to health CAs. The question arises how health CAs can be tested for their ability to understand user input formulated with different levels of health literacy. The aim of this paper is to present an approach that allows to simulate user input with different levels of health literacy and to investigate the accuracy of a health CA when confronted with this input. Specifically, we aim at answering the following research question: *How can LLMs be used to efficiently assess health CAs' ability to deal with input from a diverse set of patient populations?*

We will propose an evaluation framework that uses LLMs to emulate patient populations with varying levels of health literacy through in-context learning [5], with the aim of assessing health CAs. LLMs, defined as "deep learning models with a huge number of parameters trained in an unsupervised way on large volumes of text" [6], have demonstrated human-level performance in various academic and professional exams and benchmarks [7]. We therefore believe in their potential to effectively simulate user input.

2. Methods

2.1. Evaluation framework

Our evaluation framework consists of a fully automated procedure and a semi-automated procedure that requires the manual assessment of results by clinicians. Both procedures require the definition of QAPs. In the first approach, alternative questions are generated based on predefined questions, a task prompt and patient vignettes. The task prompt instructs the LLM what to do, i.e., to produce ten alternative questions for an existing question, based on a patient vignette. Patient vignettes are prompts that instruct an LLM to act as certain type of patient (e.g. having low (health) literacy). The generated alternative questions are sent to the CA with its answers being evaluated. Specifically, its answers are compared with the ground truth pre-defined answer. Pseudo-code for the automated evaluation is shown in Figure 1.

The total number of correctly and incorrectly classified questions is stored and used to calculate the accuracy averaged over each patient vignette and averaged over all questions, see Eq. (1). A question is considered correct if the answer returned by the

system matches the original predefined question. Questions for which an incorrect or no answer is returned are treated as misclassified.

Algorithm Automated evaluation of Question Variations

```
Input: dict predefinedContent, list patientVignettes
for each question-answer pair (question, groundTruthAnswer) in predefinedContent do
    for each vignette patientVignette in patientVignettes do
        generatedQuestionVariations ← SendToLLM(question, vignette)
        Save (generatedQuestionVariations, vignette, question, groundTruthAnswer) object
    end for
end for
for each object (generatedQuestionVariations, vignette, question, groundTruthAnswer) in the saved objects do
    for each question variation generatedQuestion in generatedQuestionVariations do
        generatedAnswer ← SendToQAmodule(generatedQuestion)
        Save (generatedQuestion, generatedAnswer) tuple to object
    end for
end for
Compute count of question-answer pairs: totalPairs ← CountAllPairs(saved objects)
Compute count of correct answers: correctCount ← CountCorrectAnswers(saved objects)
Compute percentage of correct answers: accuracy ← (correctCount / totalPairs) × 100 %
Output: accuracy
```

Figure 1. Pseudo-code of automated evaluation

$$Total\ Accuracy = \frac{Correct\ answers}{QAPs * No.\ of\ patient\ vignettes * No.\ of\ variations} \quad (1)$$

For the second, semi-automated approach, possible patient questions are generated for the domain covered by the CA. This procedure is based on a different task prompt that instructs the LLM to generate questions, impersonating each defined patient vignette. This approach additionally tests the CA's ability to handle unexpected user input. The generated questions are sent to the CA and stored together with the generated answer. To assess the quality of the CA, the generated answers are independently assessed and labelled as incorrect, partially correct, or correct by two authors (KD, DR). Inconsistencies are resolved by discussion. The evaluation score is the percentage of correct answers.

2.2. Case study

In order to test the proposed evaluation framework, we are using an existing health CA that is a rule-based, FHIR-conformant system for collecting the medical history of patients [8]. We added a QA module containing pre-defined QAPs. This module will be tested using the framework. The QA module is based on similarity matching: Let E = $\{e_1, e_2, \ldots, e_n\}$ be the set of n embedded predefined questions, where each e_i represents an embedded string. The embedded patient question to be answered is denoted as e_{query}. The cosine similarity between two embedded strings e_i and e_j is denoted as cosine similarity (e_i, e_j). To identify the most similar string, we formulate the objective as finding the index i that minimizes the cosine similarity, see Eq. (2).

$$Matching\ question = argmin_i \left(\text{cosine similarity}\left(e_{\text{query}}, e_i\right)\right) \quad (2)$$

The QA system does not provide an answer if the similarity score is too low or if the patient question is less than three words long. The minimum length of three words for questions ensures that enough information is available to be compared with the pre-defined content. Questions that cannot be answered are stored for continuous improvement and addition of QAPs. Furthermore, the system enables patients to flag wrong or complicated answers. Text embeddings are created with a pre-trained model [9] based on the sentence-transformers library [10] and stored in Chroma, an open-source vector database [11]. The QA module is currently implemented for the field of mammography. Two radiologists developed a set of 33 German QAPs for the domain of mammography, comprising popular questions asked by patients before an intervention (e.g. regarding costs, preparation, pain, expectations, duration etc.). These QAPs are integrated in the QA system.

3. Results

A total of three patient vignettes were experimentally developed, representing patients with high health literacy, low health literacy, and low literacy of German. The following text shows the patient vignette (translated from German) for low health literacy: *"You are a patient with low health literacy, and you are not well informed about the healthcare system. You use simple language and do not comprehend complicated medical terms. You will soon undergo a mammography."* For the automated evaluation, the following translated prompt was iteratively and experimentally developed to obtain ten question variations per QAP based on each patient vignette: *"Define, based on the following question, ten different variations of this question. Adapt your choice of words to your health literacy and literacy level. Return only the question, separated by a line break (\n)."* As LLM, OpenAI GPT-4 was used [7]. The results of the automated evaluation process are described in Table 1. In total, 990 alternative questions were generated by the LLM (33 QAPS * ten alternatives * three patient vignettes). Eleven questions (0.01 %) were not answered by the QA module as they contained fewer than three words, counting towards the number of errors.

Table 1. Results of case study (automated evaluation) tested on 990 alternative questions

Patient vignette	Errors (n)	Accuracy
High health literacy	101	0.69
Low health literacy	75	0.77
Low German literacy	69	0.79
Total	245	0.75

The results of the semi-automated evaluation involving manual judgement are described in Table 2. The following task prompt was used: *"Define, based on the following description [i.e. the patient vignette], ten different questions to ask your doctor. Adapt your choice of words to your health literacy and literacy level. Return only the question, separated by a line break (\n)."* In total, 30 new questions for the domain of mammography were generated (ten alternatives * three patient vignettes). Four questions were not answered by the QA module as they contained fewer than three words and no

answer was returned for three questions due to no pre-defined question being similar enough. Both categories are counting towards incorrect answers.

Table 2. Results of case study (semi-automated evaluation) tested on 30 automatically generated questions

Patient vignette	Incorrect (n)	Partial (n)	Correct (n)
High health literacy	5	5	0
Low health literacy	8	0	2
Low German literacy	5	0	5
Total	18	5	7

Obtaining the question variations and new questions resulted in total costs of $3.43 and therefore $0.10 per QAP using three patient vignettes. We make the source code of the evaluation framework including the QAPs for mammography as well as all developed prompts publicly available via Zenodo (10.5281/zenodo.10782323).

4. Discussion

Regarding automated evaluation, the results show that the 'high health literacy' patient vignette had the lowest accuracy. This may be because the LLM, as designed, generated relatively long alternative questions that tended to use more sophisticated wording, complicating the matching with possible responses. Conversely, the proportion of correctly answered questions derived from the 'low health literacy' vignette was ten percentage points higher. This might be explained by simple and short generated question variants. However, it can be assumed that patients with such a low level of German language literacy would not comprehend the pre-defined provided answers since their understanding requires a high health literacy and high readability skills.

As far as the semi-automated evaluation is concerned, none of the high literacy questions could be fully answered by the QA system. 50% of the questions were partially answered. This could be due to relatively long and complicated question variations, similar to the automated scoring. Only for the low health literacy vignette did the QA system report that it could not find a suitable answer in three cases. Although this response is technically wrong, it is a more desirable behavior than providing a completely wrong answer. In total, only seven newly generated questions were answered correctly, showing that the evaluated QA system does not yet generalize well to new data. The development of patient vignettes proved to be challenging: On the one hand, each patient vignette should lead to a unique formulation. On the other hand, the generated questions still need to be realistic. We recognized that for patient vignettes with high health literacy, the LLM tends to exaggerate the use of complex terms and phrases and domain-specific terms. Nevertheless, the QA system was able to answer 69 % of these questions correctly in the automated evaluation. See Table 3 for an example of an original question and LLM-generated question using the vignette-based alternatives.

Table 3. Comparison of original and LLM-generated questions

Group	Variant (translated from German)
Original question	How long does a mammogram take?
High health literacy	Assuming I undergo a mammogram, how long would I be expected to spend in the radiology department?
Low health literacy	How long will I sit there when they do this mammogram?
Low German Literacy	Mammogram, much time needed?

To improve performance of our QA module, the following improvement strategies can be applied: If two questions with different answers are very similar to each other, the model may constantly misclassify one of them. To improve this, these questions could either be merged into one QAP, or each question could be rephrased. Next, the distance threshold defines the minimum level of similarity between a predefined question and a patient question for the system to return an answer. As every question exceeded this threshold in the automated test, a higher threshold could lead to higher precision, albeit at the expense of accuracy. Finally, better models are continually being published. The model used to implement the evaluated QA module is easily replaceable due to its compatibility with the sentence-transformers framework.

The limitations of our study are as follows: Only one commercially available LLM was tested. In future, the performance of other LLMs, including open-source models, should be compared. Moreover, our case study only focused on a single clinical domain with a relatively small register of QAPs (n = 33) and newly generated questions (n = 30).

In this paper, we present a scalable, semi-automated evaluation framework to evaluate domain-specific QA systems. The developed task prompts and patient vignettes can be easily used with other models. Using LLMs and patient vignettes, QA inputs from different patient populations can be easily simulated. Based on this evaluation framework, QA systems can be evaluated with low effort cost-effective, fostering their application in clinical practice. Future research directions include adapting this evaluation framework to QA systems without pre-defined content and elaborating on the development of more detailed prompts to simulate patient populations. Furthermore, we recommend investigating whether LLM-based adaptation of responses based on the patient's (health) literacy level improves understandability and patient satisfaction.

References

[1] K. Denecke, 'Potential and pitfalls of conversational agents in health care', Nat Rev Dis Primers, vol. 9, no. 1, Art. no. 1, Nov. 2023, doi: 10.1038/s41572-023-00482-x.

[2] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, 'Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering', Transactions of the Association for Computational Linguistics, vol. 11, pp. 1–17, Jan. 2023, doi: 10.1162/tacl_a_00530.

[3] K. Singhal et al., 'Large language models encode clinical knowledge', Nature, vol. 620, no. 7972, pp. 172–180, Aug. 2023, doi: 10.1038/s41586-023-06291-2.

[4] C. Liu et al., 'What is the meaning of health literacy? A systematic review and qualitative synthesis', Fam Med Community Health, vol. 8, no. 2, p. e000351, May 2020, doi: 10.1136/fmch-2020-000351.

[5] T. Brown et al., 'Language Models are Few-Shot Learners', in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2020, pp. 1877–1901.

[6] A. Birhane, A. Kasirzadeh, D. Leslie, and S. Wachter, 'Science in the age of large language models', Nat Rev Phys, vol. 5, no. 5, Art. no. 5, May 2023, doi: 10.1038/s42254-023-00581-4.

[7] OpenAI et al., 'GPT-4 Technical Report'. arXiv, Dec. 18, 2023. doi: 10.48550/arXiv.2303.08774.

[8] K. Denecke, N. Cihoric, and D. Reichenpfader, 'Designing a Digital Medical Interview Assistant for Radiology', Stud Health Technol Inform, vol. 301, pp. 60–66, May 2023, doi: 10.3233/SHTI230012.

[9] P. May and deepset GmbH, 'German BERT large paraphrase cosine'. Deutsche Telekom, Dec. 17, 2023. Accessed: Jan. 05, 2024. [Online]. Available: https://huggingface.co/deutsche-telekom/gbert-large-paraphrase-cosine

[10] N. Reimers and I. Gurevych, 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'. arXiv, Aug. 27, 2019.

[11] Chroma, 'Chroma'. Chroma, Jan. 05, 2024. Accessed: Jan. 05, 2024. [Online]. Available: https://github.com/chroma-core/chroma