Leipzig University
Institute of Computer Science
Degree Programme Data Science M.Sc.

Implicit Evaluation of Health Answers from Large Language Models

Master's Thesis

Jonas Probst
Born Sept 10, 1995 in Stuttgart

Matriculation Number 3466651

1. Referee: Prof. Dr. Martin Potthast
2. Referee: Dr. Harrisen Scells

Submission date: January 16, 2024

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, January 16, 2024

. . . .

Jonas Probst

Abstract

With the release of ChatGPT, open-ended generation of text became the biggest use case of Large Language Models (LLMs). Meanwhile, LLM evaluation focuses on classical NLP tasks like single-choice question answering or text classification, which do not represent the LLMs' capabilities in long-form question answering (LFQA). The lack of evaluation of open-ended questions is especially concerning in the medical domain, as answers that are misleading or wrong could have significant impact on the users personal health. Using human experts to compare generated answers is considered the gold standard in this space, but it leads to high costs and slower evaluation procedures, while also introducing subjectivity to the evaluations. In this thesis, we present a retrieval-based implicit evaluation method for LFQA, aiming to make the evaluation process faster, cheaper and more repeatable. Using a dataset of queries and associated documents, which were evaluated by human annotators, we first compare multiple retrieval methods for retrieving documents that are relevant, readable, and credible. We then use the most effective retrieval method to rank new answers generated by LLMs against the web answers from the dataset. Because the retrieval method is previously evaluated to produce rankings similar to the human evaluations, we assume that it ranks the generated LLM answer close to where a human would rank it. Our findings demonstrate that the proposed retrieval-based implicit evaluation ranks the effectiveness of various LLMs in a similar order as other benchmarks, underscoring the validity of the approach. Additionally, we show that the LLM ranking improves with model size and with more sophisticated prompting strategies, which aligns with trends observed in the literature. This work is a first step towards building a more automated evaluation framework for LFQA, which could decrease development costs and ensure comparability of different LLMs, even if they are not evaluated by the same research group. Because the most effective model in our research (ChatGPT) already ranks best on nearly every query, we encourage future research to produce a more challenging dataset, enabling the comparison of more advanced models.

Contents

Chapter 1

Introduction

After the release of ChatGPT[1] by OpenAI in 2022 a number of other chatbots like Anthropic's Claude[2] or Google's Bard[3] were published in the following months. Especially ChatGPT became successful very quickly, gaining 100 million active users in the first year after release, according to the OpenAI DevDay Opening Keynote[4]. But the chatbots by other companies also got regular updates and a lot of attention from media and the companies, indicating high public interest in this space. Based on the findings by De Choudhury et al. (2014), who show that a large amount of users turns to search engines or social media sites for medical information, we can assume that a large portion of users turns to ChatGPT and other chatbots with similar medical questions. This assumption is supported by studies investigating possible use cases of ChatGPT in the medical domain, which mention the use of ChatGPT by doctors, nursers or medical students as one possible application (Dave et al. (2023), Khan et al. (2023)). While multiple benchmarks for answering questions in a single choice format or with short, factual answers are usually reported in the technical reports accompanying the release of new models, it is much harder to evaluate the LLMs on open-ended questions that are common in the medical domain.

As Ouyang et al. (2022) show, the current NLP datasets and benchmarks that are usually applied to evaluate LLMs do not reflect the described use case of answering open-ended questions. According to their data, only 18% of GPT-3 API calls are targeted at conventional NLP tasks like classification or single choice QA tasks. On the other hand, 57% of API requests lead to open-ended generation. Considering that this data originates from the GPT-

[1] `https://chat.openai.com/`

[2] `https://claude.ai/`

[3] `https://bard.google.com/chat`

[4] `https://www.youtube.com/watch?v=U9mJuUkhUzk`

3 API and not the conversation-focused ChatGPT model, it is reasonable to assume that open-ended generation now makes up even more of the requests. This shows that traditional QA tasks like single or multiple choice QA or extracting answers from a given text are not representative of the task the chatbots are supposed to perform.

The relatively new field of long-form question answering (LFQA) deals with the answering of open-ended questions by automated systems. As shown by Xu et al. (2023), the current automated methods of evaluating those LFQA systems are still lacking compared to manual human evaluation, which is often performed by crowd workers. However, Xu et al. (2023) also mention multiple drawbacks of the human evaluation approach. For one, annotators need to be well-trained and have a solid foundational knowledge of the question field. This level of expertise is usually lacking among crowd workers. Furthermore, the answer length makes evaluating long-form QA much more demanding than simple QA tasks (Krishna et al. (2021)), increasing the duration it takes to process one sample. Together, these problems lead to an increased need for well-trained human annotators, which are hard to find and expensive. Additionally, human evaluations for larger datasets are time-consuming, slowing down the development cycle particularly in the fine-tuning of LLMs.

Krishna et al. (2021) demonstrate a high rate of disagreement among annotators when choosing between two given answers to the same question. They attribute this to the difficulty of judging answer quality, due to the many factors that make up a good answer. The subjectivity of the evaluation also poses problems for creating more widely used LFQA benchmarks based on human evaluation. Even if the questions for all benchmarked models are the same, the subjectivity introduced by human annotators between the different models renders the resulting scores incomparable, assuming not every question is annotated by the same crowd worker for each LLM.

In addition to making the evaluation of correctness more difficult, open-ended generation of text also necessitates that the quality of the answers is evaluated in a multidimensional manner. In contrast to other QA tasks, LFQA has no simple notion of a good answer since answer quality is not only determined by the correctness of the answer, but also by its readability, relevance and credibility. Current evaluation methods like accuracy for single choice question answering or the overlap between the extracted answer and the ground truth for extractive question answering are not able to capture the multidimensional nature of answer quality. To reduce human evaluation costs and enable large-scale, multidimensional evaluation and comparison of chatbots, automated methods are necessary.

In this thesis, we propose a new evaluation method for long-form question answering based on information retrieval techniques. Using a dataset based on health questions from Goeuriot et al. (2021), we construct a benchmark consisting of multiple queries and human-generated answers based on web content. We then use retrieval methods to rank answers generated by different LLMs against those human-generated web documents. Based on the rank of the generated answer, we can compare the quality of the answers by different LLMs. This method captures the multidimensional nature of answer quality, assuming the retrieval method is effectively ranking documents in terms of relevance, readability and credibility. Evaluating the multidimensional retrieval performance of different retrieval methods is therefore an important part of this thesis.

With the answer evaluation process being automated, this method allows for large-scale evaluation of LLMs, including the evaluation of different prompting strategies, the assessment of answer consistency across a model, and the analysis of how answer quality varies with the number of model parameters. This reduction in human evaluations would consequently lower the costs associated with developing and fine-tuning LLMs for LFQA tasks. It would also lead to more consistency in evaluating generated answers, limiting the subjective component brought in by human annotators.

1.1 Research Questions

In order to evaluate the proposed retrieval-based implicit evaluation method, we formulate one overarching research question:

Is a retrieval-based implicit evaluation method using human-written web content a viable approach for assessing the quality of health answers generated by LLMs?

To investigate the main research question, we formulate two supporting research questions:

- **RQ1:** Which factors influence the effectiveness of LLMs on the proposed evaluation method?
- **RQ2:** How does the effectiveness of LLMs on existing benchmarks relate to their effectiveness on the proposed retrieval-based implicit evaluation method?

1.2 Scope and Limitations

This thesis is intended as a first step in investigating the use of information retrieval techniques for evaluating the LFQA capabilities of LLMs. It is not meant to be a comprehensive evaluation of the proposed method, but rather a proof of concept on which future work can build.

The benchmark dataset is not built from scratch, but is based on the dataset from Goeuriot et al. (2021), which is not originally constructed for evaluating LFQA. Our preprocessing of the web documents still leaves some noise, indicating that more sophisticated methods for extracting the core message from the web documents could yield a more challenging dataset. The dataset is therefore not perfectly suited for the task, but is still useful for a first evaluation of the proposed method.

Additionally, the dataset is only contains queries and documents in the English language, so the capabilities of the models in other languages are not considered.

The domain of the datasets is restricted to medical queries; other fields are not explored in this work.

1.3 Structure of the Thesis

Following this Introduction, in Chapter **??** we delve into the related work. First, the general field of evaluating Large Language Model in the context of question answering is introduced (Section **??**) This includes an overview of different question answering tasks and their evaluation methods.

Subsequently, in Section **??** the different retrieval methods used in this work are presented, alongside the evaluation metrics used to compare them. This section is concluded with a discussion on how retrieval methods have previously been used to evaluate NLP tasks.

Chapter **??** describes the experimental setting, from dataset collection and preparation over development of the different retrieval pipelines and generation of LLM responses to the questions in the dataset.

The experimental results are presented in Chapter **??**, aiming to provide the necessary groundwork for discussing the research questions in the following Chapter **??**, which is done in the section following the results.

We conclude this thesis with Chapter **??** on the outlook on future work and a conclusion.

Bibliography

Dave, T., Athaluri, S. A., and Singh, S. (2023). Chatgpt in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Frontiers in Artificial Intelligence*, 6:1169595.

De Choudhury, M., Morris, M. R., and White, R. W. (2014). Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1365–1376.

Goeuriot, L., Suominen, H., Pasi, G., Bassani, E., Brew-Sam, N., Sáez, G. N. G., Kelly, L., Mulhem, P., Seneviratne, S., Upadhyay, R., Viviani, M., and Xu, C. (2021). Consumer Health Search at CLEF eHealth 2021.

Khan, R. A., Jawaid, M., Khan, A. R., and Sajjad, M. (2023). Chatgpt-reshaping medical education and clinical management. *Pakistan Journal of Medical Sciences*, 39(2):605.

Krishna, K., Roy, A., and Iyyer, M. (2021). Hurdles to progress in long-form question answering. *arXiv preprint arXiv:2103.06332*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Xu, F., Song, Y., Iyyer, M., and Choi, E. (2023). A critical evaluation of evaluations for long-form question answering. *arXiv preprint arXiv:2305.18201*.