Ranking Generated Answers

On the Agreement of Retrieval Models with Humans on Consumer Health Questions

Sebastian Heineking[1], Jonas Probst[1], Daniel Steinbach[3], Martin Potthast[2], and Harrisen Scells[1]

[1] Leipzig University
[2] University of Kassel, hessian.AI, and ScaDS.AI
[3] University of Leipzig Medical Center

Abstract. Evaluating the output of generative large language models (LLMs) is challenging and difficult to scale. Most evaluations of LLMs focus on tasks such as single-choice question-answering or text classification. These tasks are not suitable for assessing open-ended question-answering capabilities, which are critical in domains where expertise is required, such as health, and where misleading or incorrect answers can have a significant impact on a user's health. Using human experts to evaluate the quality of LLM answers is generally considered the gold standard, but expert annotation is costly and slow. We present a method for evaluating LLM answers that uses ranking signals as a substitute for explicit relevance judgements. Our scoring method correlates with the preferences of human experts. We validate it by investigating the well-known fact that the quality of generated answers improves with the size of the model as well as with more sophisticated prompting strategies.

Keywords: evaluation, consumer health search, large language models

1 Introduction and Related Work

Search engines are used to ask questions in domains where a wrong answer can mean a high risk for the user [15]. Examples include health, medicine, finance, and law. Due to the many factors that influence answers in these domains, open-ended questions—as opposed to single-choice or factual questions—are more common here. The use of LLMs for retrieval-augmented generation (RAG) and as chatbots in conversational search engines, as well as the tendency of some models to confirm user bias [2, 8, 18] call for a careful, large-scale evaluation of question-answering systems that require domain expertise. In fact, this requirement may go beyond expert domains: Ouyang et al. [13] found that only 18% of GPT-3 API calls are single-choice questions or text classification queries, while 57% are open-ended questions.

While benchmarks with single-choice or factual questions allow a comparably easy automatic evaluation of LLMs [4, 19], their quality measures, such as text overlap or text similarity, are not suited for evaluating open-ended questions and their complex, nuanced answers. Here, automatic evaluation cannot compete with manual evaluation by humans [19]. As a remedy to achieve at least semi-automation, crowdworkers are often employed for evaluation, but they require extensive training and knowledge of the domain in question. In this context, Krishna et al. [9] note that the evaluation of answers to open-ended questions is much more challenging than that of answers to single-choice questions due to their length. Longer answers increase the time needed to process an example, leading to low annotator agreement when choosing between two answers to the same question. Another factor contributing to disagreement among annotators is that the quality of answers often cannot be reduced to a single dimension but depends on several factors [9], requiring a multifaceted evaluation. Sakai [16] proposed to evaluate the quality of conversational systems based on criteria such as fluency, soundness, and explainability. The proposed criteria, however, currently require a human-in-the-loop evaluation, and automated, scalable solutions remain an open research question. Farzi and Dietz [5] proposed to evaluate LLM answers using a grading rubric of questions that answers must address. However, the method is focused only on relevance and other criteria are not accounted for.

We propose a new evaluation method for generated answers to bridge the gap between high-quality human assessments and more efficient automatic approaches. Our method uses ranking signals from annotated corpora to measure the effectiveness of generated answers. For this method, we took inspiration from a ranking-based evaluation method for machine translation, where the translations generated by different models are ranked together with reference translations to compare the models' effectiveness [3, 7, 10]. To our knowledge, ranking-based evaluation has yet to be used to evaluate other tasks. Since manual annotations are not required for each new answer, this method allows for scalable evaluation of LLMs, including comparing different prompting strategies and model sizes. Furthermore, it facilitates consistent evaluation of generated answers by limiting the room for subjective judgements to the initial annotations. In our investigation, we focus on the consumer health domain and analyze the influences of factors such as model size and prompting strategy on the effectiveness of LLMs for rank-based evaluation. We further conduct a user study with a healthcare professional to validate our ranking method in comparison to how they rank answers from LLMs, as suggested by Arabzadeh and Clarke [1].[4]

2 NRP: Normalized Rank Position for Generated Answers

We propose normalized rank position (NRP), an automatic method to evaluate answers generated by LLMs. It requires a set of queries and documents with query relevance judgments. First, each LLM to be evaluated is prompted with each query. Second, each generated answer is independently ranked along with the documents, using a retrieval model known to have a high effectiveness in ranking the documents with respect to their judgments. Third, the effectiveness of each LLM is estimated based on the ranks of its responses.

[4] Code and data: https://github.com/.

We formalize this approach as follows:

1. Collect a set of queries $Q = \{q_1, \ldots, q_n\}$ and a set of human-written documents $D_Q = \bigcup_{q \in Q} D_q$, where $D_q = \{d_1, \ldots d_m\}$ contains documents that are annotated with respect to relevance to q, or other quality dimensions.
2. Evaluate a set of retrieval models $\mathcal{R}$ on D across the quality dimensions.
3. Use a set of LLMs $M \in \mathcal{M}$ to generate answers $a_{l,i}$ for all queries $q_i \in Q$.
4. Add a generated answer $a_{l,i}$ to the set of documents D_{q_i} for query q_i and rank the set using the best model from $\mathcal{R}$. The rank position of the answer is an indicator of the effectiveness of the LLM M for the query.

We evaluate the LLM answer effectiveness using a measure we call normalized rank position (NRP). NRP is the absolute position r of the answer in the ranking, divided by the number of documents for the query $\text{NRP} = 1 - \frac{r}{|D_{q_i}|}$.

3 Experimental Setup

Data Collection and Preparation. For evaluation, we use the CLEF eHealth 2021 test data [6]. It comprises 55 health-related queries obtained from medical experts and Reddit discussion and relevance judgments for Web documents as well as Reddit and Twitter posts. With API changes at Reddit and Twitter, the content from these platforms has become unavailable. We therefore obtained only the original Web documents from CommonCrawl, discarded those containing fewer than 50 characters in the HTML body, and extracted plain text using the Resiliparse library.[5] We were able to restore 6,692 Web documents with judgments, omitting 5 queries exclusively paired with Twitter and Reddit posts.

Retrieval Pipeline. For ranking, we used two lexical retrieval models, TF-IDF and DPH, and four transformer-based models: `ColBERTv1`, `ColBERTv2`, `monoT5`, and `duoT5`. Except for `ColBERTv2`, the versions available via PyTerrier [11] were used. For `ColBERTv2`, the implementation is the one provided by the authors of the model.[6] All transformer-based models are fine-tuned on MSMARCO. The most effective model in terms of nDCG@10 across the three quality dimensions of relevance, readability, and credibility is `monoT5` with scores of 0.645, 0.813, and 0.722, respectively. Hence, `monoT5` is used in all further experiments.

Generating Responses. To generate answers, we selected LLMs that differ in (1) number of parameters, (2) amount and type of training data, and (3) pre-training and fine-tuning. We use the base, medium, large, and XL variants of `GPT-2`, the instruction-tuned `Falcon` 7B and `LLaMA-2` 7B and 13B, and `GPT-3.5-turbo-0613`. Across all models, we fixed the maximum number of new tokens to 512, the temperature to 0.75, top-k to 50, top-p to 0.95, and the repetition penalty to 1.2. For `ChatGPT`, we could only set the temperature and maximum

[5] https://resiliparse.chatnoir.eu/en/stable/

[6] https://github.com/stanford-futuredata/ColBERT

number of new tokens. Prior research has shown the prompt formulation can greatly affect answer quality [14]. Therefore, after a pilot study, we included multiple prompts in our experiments for comparison:

No Prompt: *query*
Short QA Prompt: Q: *query* A:
Long QA Prompt: Question: *query* Answer:
MultiMedQA Prompt [12]: You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically grounded answers to common consumer search queries about health. Question: *query* Complete Answer:

We sample ten responses for each query and LLM to reduce the effect of random variations in generated answers.

User Study. Finally, we measured how well NRP reflects the effectiveness of generated answers compared to an expert annotator (a medical doctor and co-author of this paper). The annotation task involves ordering responses to a query from most relevant, readable, and credible to least. To keep the annotation task feasible, we sampled 20 topics and chose five answers for each topic for annotation: the top-ranked responses by `ChatGPT`, `Llama-2 13B`, `GPT-2 XL`, `GPT-2`, and the top-ranked relevant Web page (as per `monoT5`). The generated answers were obtained using the MultiMedQA prompt and agreement between the expert and the ranking model was measured using RBO [17] ($p = 1$) and Kendall's τ.

4 Results & Discussion

Factors Influencing LLM Answer Quality. We investigate the effect of prompting strategy and model size on NRP. In total, we rank 16,000 generated answers: Each of the 8 models generated 10 answers for 4 different prompting strategies and 50 queries. Figure 1 shows the NRP scores for each LLM, grouped by prompt. As illustrated in the figure, the choice of prompting strategy influences the position of generated answers, especially for models with more parameters. When no context is given to the model ("No Prompt"), the effectiveness decreases for many models, with the exception of `ChatGPT`. Many small models are able to achieve a higher effectiveness with the MultiMedQA prompt than the next larger model with no prompt. This gain in effectiveness emphasizes the importance of prompting, even for models like `Falcon` and `Llama-2` that are already fine-tuned for instruction tasks. Of all models, `ChatGPT` is the most effective with a mean NRP of 0.998. `Falcon 7B` is the least effective of the fine-tuned models with the `Llama` models being slightly more effective. The `GPT-2` variants show the lowest effectiveness overall, but improve as the number of parameters increases. Figure 2 visualises this trend by comparing the number of parameters to NRP scores when using the MultiMedQA prompt. The largest improvement in NRP is between `GPT-2 Medium` and `GPT-2 Large`, and between `GPT-2 XL` and `Falcon 7B`/`Llama-2 7B`. For the small to medium-sized fine-tuned models, specifically the 7B variant of `Falcon` and the 7B and 13B variants of `Llama-2`,

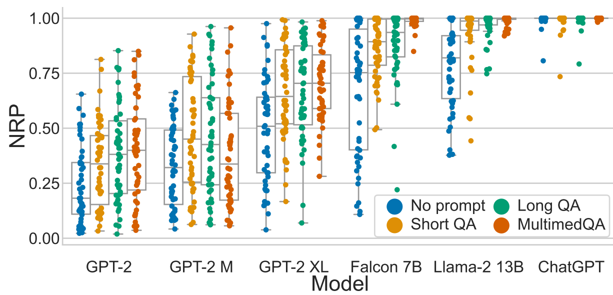

Fig. 1. NRP of LLM answers, averaged over the ten generated answers per question, grouped by prompt. Points are single answers. `GPT-2 L` and `Llama-2 7B` not shown.

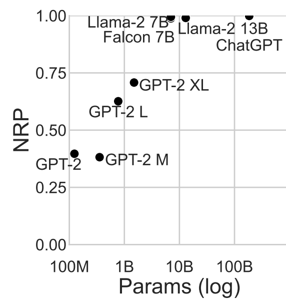

Fig. 2. Number of model parameters vs. NRP using the MultiMedQA prompt.

the trend is less clear. `Llama-2 7B` is more effective than the 7B variant of `Falcon`, achieving an NRP of 0.995 compared to 0.988 by `Falcon 7B`. While they have the same number of parameters, the observed differences could be explained by differences in training data and the more sophisticated fine-tuning strategy of `Llama-2`. More surprisingly, `Llama-2 7B` on average ranks slightly better than the larger variant, which has nearly twice as many parameters. The similar rankings could indicate a saturation of the `Llama-2` model's effectiveness around that parameter count for our specific task. To summarize, prompting strategy and model size have a large influence on the effectiveness of generated answers and these effects are measurable with NRP.

Agreement with Human Expert Preferences. Finally, we compare the preferences of our chosen ranking model with that of an expert annotator. Figure 3 visualizes this comparison by showing where documents are ranked across the 20 topics. Each column in the figure corresponds to an answer from an LLM, or a web document (denoted 'Document') at that rank position. The flow between columns indicates the source of the next answer in the rankings where a given model has achieved the respective higher ranking. For example, in Figure 3 (top), `Llama-2 13B` and `ChatGPT` are the only two models to have answers in the first position across the 20 rankings. The next answer in rankings where `Llama-2 13B` is on first rank comes mostly from `ChatGPT`, except for one topic where the answer from `GPT-2 XL` was ranked in second position. Comparing the top and bottom figures, only minor differences between the ranking model and the expert are observed. Both figures reveal a common trend: `ChatGPT` is the most preferred answer, `Llama-2 13B` is the second, `GPT-2 XL` and the web document are often tied for Positions 3 and 4. `GPT-2`, finally, is the least preferred answer, occurring almost always in the last position (and only ever higher in the second-last position). An average RBO of 0.84 and a Kendall's τ of 0.64 underlines the high agreement between our chosen ranking model and the judgments by the expert annotator.

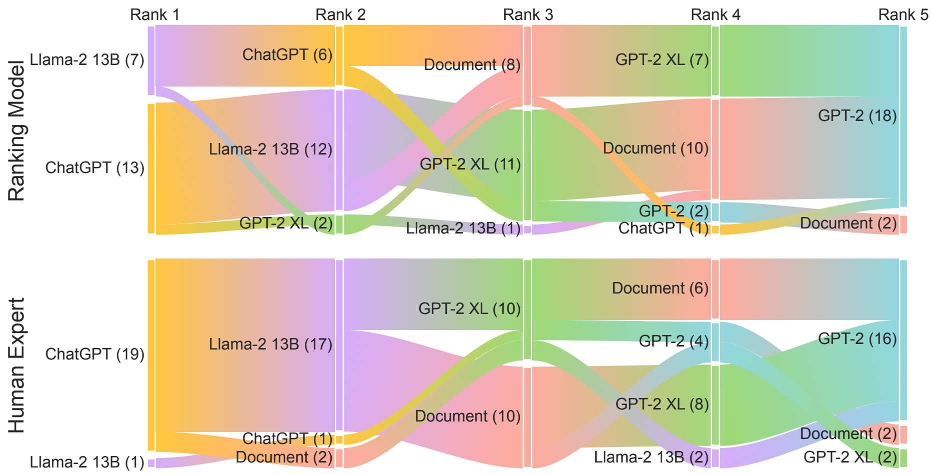

Fig. 3. How our chosen ranking model ranks answers (top) versus how the expert annotator ranks answers (bottom). Each column is a rank position. Numbers in brackets are the amount of answers of that type at that rank position, e.g., GPT-2 (16) indicates 16 answers at that rank position come from `GPT-2`.

5 Conclusions

We developed a method to scalably and automatically evaluate generated answers to consumer health questions. Our NRP score evaluates model effectiveness by ranking answers in comparison with human-written documents. NRP does not rely on 'gold standard' answers, but uses existing texts. We investigated the factors that led to higher effectiveness when using the measure, and whether human evaluations of generated answers agree with the rankings from an effective ranking model. However, the results of our experiments may be limited in their ability to discriminate highly effective LLMs as the dataset we used was potentially too easy for these models. We leave this investigation for future work. We demonstrate that state-of-the-art ranking functions, once validated using a test collection with multi-dimensional relevance assessments, can be used to effectively discriminate high-quality from low-quality responses. The ranking functions can even be used to discriminate between responses from similar LLMs. Furthermore, once the test collection and ranking function are finalized, other LLMs can be evaluated using the setup in an entirely offline and scaleable manner. The experiments in this paper focus on evaluating generated answers for consumer health search questions. For future work, we will develop and adapt additional test collections for other domains. Another avenue for future research is the extension of NRP to assess retrieval-augmented generation (RAG) systems that ground their answers in documents by referring to them. This paper contributes a new way to automatically assess LLM capabilities by evaluating them in the context of human-annotated documents.

References

[1] Arabzadeh, N., Clarke, C.L.A.: A comparison of methods for evaluating generative IR. CoRR **abs/2404.04044** (2024), https://doi.org/10.48550/ARXIV.2404.04044, URL https://doi.org/10.48550/arXiv.2404.04044

[2] Berberette, E., Hutchins, J., Sadovnik, A.: Redefining "hallucination" in llms: Towards a psychology-informed framework for mitigating misinformation. ArXiv **abs/2402.01769** (2024)

[3] Duh, K.: Ranking vs. regression in machine translation evaluation. In: Proceedings of the Third Workshop on Statistical Machine Translation, p. 191?194, StatMT '08, Association for Computational Linguistics, USA (2008), ISBN 9781932432091

[4] Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., Auli, M.: Eli5: Long form question answering. In: ACL, pp. 3558–3567 (2019)

[5] Farzi, N., Dietz, L.: Pencils down! automatic rubric-based evaluation of retrieve/generate systems. In: Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 175–184, ICTIR '24, Association for Computing Machinery, New York, NY, USA (2024), ISBN 9798400706813, https://doi.org/10.1145/3664190.3672511, URL https://doi.org/10.1145/3664190.3672511

[6] 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., Xu, C.: Consumer Health Search at CLEF eHealth 2021. In: (CEUR-WS.org), C.W.P. (ed.) Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, Bucharest, Romania (2021)

[7] Guzmán, F., Joty, S., Màrquez, L., Nakov, P.: Pairwise neural machine translation evaluation. In: Zong, C., Strube, M. (eds.) Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), vol. 1, pp. 805–814, Association for Computational Linguistics, Beijing, China (Jul 2015), https://doi.org/10.3115/v1/P15-1078

[8] Koopman, B., Zuccon, G.: Dr ChatGPT tell me what I want to hear: How different prompts impact health answer correctness. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 15012–15022, Association for Computational Linguistics, Singapore (Dec 2023), https://doi.org/10.18653/v1/2023.emnlp-main.928

[9] Krishna, K., Roy, A., Iyyer, M.: Hurdles to progress in long-form question answering. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., Zhou, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech-

nologies, pp. 4940–4957, Association for Computational Linguistics, Online (Jun 2021), https://doi.org/10.18653/v1/2021.naacl-main.393

[10] Li, M., Jiang, A., Wang, M.: Listwise approach to learning to rank for automatic evaluation of machine translation. In: Way, A., Sima'an, Khalil and-Forcada, M.L. (eds.) Proceedings of Machine Translation Summit XIV: Papers, Nice, France (Sep 2013)

[11] Macdonald, C., Tonellotto, N.: Declarative experimentation in information retrieval using pyterrier. In: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, pp. 161–168, ICTIR '20, Association for Computing Machinery, New York, NY, USA (2020), ISBN 9781450380676, https://doi.org/10.1145/3409256.3409829

[12] Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of GPT-4 on Medical Challenge Problems (Apr 2023)

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

[14] Reynolds, L., McDonell, K.: Prompt programming for large language models: Beyond the few-shot paradigm. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, CHI EA '21, Association for Computing Machinery, New York, NY, USA (2021), ISBN 9781450380959, https://doi.org/10.1145/3411763.3451760

[15] Roegiest, A., Pinkosova, Z.: Generative information systems are great if you can read. In: Clough, P.D., Harvey, M., Hopfgartner, F. (eds.) Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval, CHIIR 2024, Sheffield, United Kingdom, March 10-14, 2024, pp. 165–177, ACM (2024), https://doi.org/10.1145/3627508.3638345, URL https://doi.org/10.1145/3627508.3638345

[16] Sakai, T.: Swan: A generic framework for auditing textual conversational systems. ArXiv **abs/2305.08290** (2023)

[17] Webber, W., Moffat, A., Zobel, J.: A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. **28**(4) (nov 2010), ISSN 1046-8188, https://doi.org/10.1145/1852102.1852106

[18] Wester, J., Lee, M., van Berkel, N.: Moral transparency as a mitigator of moral bias in conversational user interfaces. In: Proceedings of the 5th International Conference on Conversational User Interfaces, CUI '23, Association for Computing Machinery, New York, NY, USA (2023), ISBN 9798400700149, https://doi.org/10.1145/3571884.3603752

[19] Xu, F., Song, Y., Iyyer, M., Choi, E.: A critical evaluation of evaluations for long-form question answering. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 3225–3245, Association for Computational Linguistics, Toronto, Canada (Jul 2023), https://doi.org/10.18653/v1/2023.acl-long.181"