Ranking Generated Answers

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

Sebastian Heineking[1] ⓘ, Jonas Probst[1], Daniel Steinbach[4] ⓘ,
Martin Potthast[2,3] ⓘ, and Harrisen Scells[1,2] ⓘ

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

Abstract. Evaluating the output of generative large language models (LLMs) is challenging and difficult to scale. Many 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. One such domain is health, where misleading or incorrect answers can have a negative impact on a user's well-being. 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 models trained on annotated document collections as a substitute for explicit relevance judgements and apply it to the CLEF 2021 eHealth dataset. In a user study, our method correlates with the preferences of a human expert (Kendall's $\tau = 0.64$). It is also consistent with previous findings in that the quality of generated answers improves with the size of the model and more sophisticated prompting strategies.[1]

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 [19]. 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, 10, 22] 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. [16] found that 57% of GPT-3 API calls are open-ended questions.

---

[1] Code and data: https://github.com/webis-de/ECIR-25.

While benchmarks with single-choice or factual questions allow a comparably simple automatic evaluation of LLMs [4, 23], 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 [23]. As a remedy, crowdworkers are often employed for evaluation, but they require extensive training and knowledge of the domain in question. In this context, Krishna et al. [11] 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 [11], requiring a multifaceted evaluation. Sakai [20] and Gienapp et al. [6] propose to evaluate the quality of conversational systems based on criteria such as fluency, soundness, and explainability. These criteria, however, currently require a human-in-the-loop evaluation, while automated, scalable solutions remain an open research question. Farzi and Dietz [5] propose to evaluate LLM answers using a grading rubric of questions that answers must address. The method, however, is focused only on relevance, while 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 take 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, 8, 12]. 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 a consistent evaluation of generated answers by limiting the room for subjective judgements to the initial annotations. We conduct experiments on the CLEF 2021 eHealth dataset [7] and analyze how model size and prompting strategy influence the effectiveness of LLMs as measured by our rank-based evaluation. We further validate our ranking method in a user study by correlating its results with the ranking of a healthcare professional, as suggested by Arabzadeh and Clarke [1].

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 Q and a set of documents D with expert judgements for the queries. These judgements can express the document's relevance to the query or other quality dimensions. First, a set of retrieval models is evaluated on ranking documents in accordance with the

expert judgements. Second, a set of LLMs is prompted to generate an answer a_q for each query q in the dataset. Third, the most effective retrieval model is used to rank a_q alongside all documents $D_q \subseteq D$ with expert judgements for q. The effectiveness of an LLM for q is evaluated based on the rank position of the answer it generated. The normalized rank position (NRP) is calculated based on the absolute rank position r of a_q and the number of documents in D_q:

$$\text{NRP} = 1 - \frac{r}{|D_q|}, \quad \text{where} \quad D_q = \{d \in D|\ d \text{ has expert judgements for } q\}.$$

For NRP scores to be in the range $[0, 1]$, the rank positions need to be based on zero-based numbering, so that the highest rank position is 0 instead of 1.

If the expert judgements encompass multiple quality dimensions, different retrieval models might be the most effective at ranking documents in accordance with a given dimension. In these cases, a dimension-specific NRP can be calculated using the most effective retrieval model for each dimension. The individual scores are then combined in a weighted sum that corresponds to the importance of each quality dimension for the given use case. For our experiments, we report a single NRP scores as the same retrieval model was the most effective on all quality dimensions.

3   Experimental Setup

Data Collection and Preparation. We used the CLEF eHealth 2021 dataset [7] for evaluation. It comprises 55 health-related queries, obtained from medical experts and social media discussions, with judgements for web documents as well as Reddit and Twitter posts. The judgements reflect the quality dimensions relevance, readability, and credibility. 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.[2] We were able to restore 6,692 web documents with judgements, 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` [9], `ColBERTv2` [21], `monoT5` [14], and `duoT5` [17]. Except for `ColBERTv2`, the versions available via PyTerrier [13] were used. For `ColBERTv2`, the implementation is the one provided by the authors of the model.[3] 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 and we calculate a single NRP based on its rankings.

---

[2] https://resiliparse.chatnoir.eu/en/stable/
[3] https://github.com/stanford-futuredata/ColBERT

Generating Answers. To generate answers, we selected LLMs that differ in (1) number of parameters, (2) amount and type of training data, and (3) pretraining and fine-tuning. We used the base, medium, large, and XL variants of `GPT-2`, the instruction-tuned `Falcon` 7B, `LLaMA-2` 7B, and 13B, as well as `ChatGPT` based on `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 number of new tokens. Prior research has shown the prompt formulation to strongly affect answer quality [18]. 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 [15]: 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 generated ten answers for each query and LLM to reduce the effect of random variations in generated answers.

User Study. Finally, we validated NRP by correlating its effectiveness estimation of different LLMs with annotations by an expert (a medical doctor and co-author of this paper). To keep the annotation task feasible, we sampled 20 of the 50 queries and provided five options per query. For each query, the expert received the most relevant web document according to `monoT5` and the top-ranked answer by `ChatGPT`, `Llama-2 13B`, `GPT-2 XL`, and `GPT-2`. They then created a joint ranking for all three quality dimensions from most relevant, readable, and credible to least. Answers were generated using the MultiMedQA prompt and agreement between the expert and `monoT5` was measured using Kendall's τ.

4   Results

Factors Influencing LLM Answer Quality. We investigated the effect of prompting strategy and model size on NRP. In total, we ranked 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 underlines the importance of prompting and is particularly visible for the instruction-tuned models `Falcon` and `Llama-2`. 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
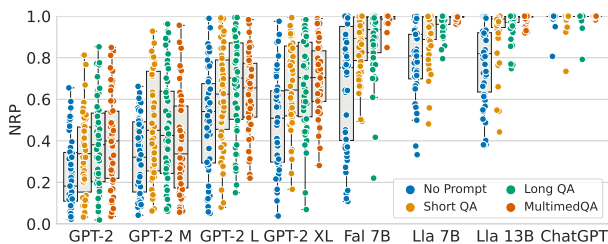
Fig. 1. NRP of LLM answers, averaged over the ten generated answers per question, grouped by prompt. Points are single answers. `GPT-2 L` and `Llama-2 7B` not shown.
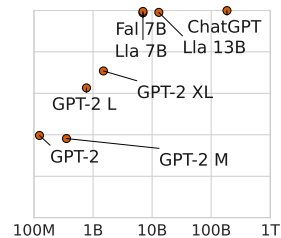
Fig. 2. Number of model parameters vs. NRP using the MultiMedQA prompt.

with the `Llama` models being slightly more effective. The `GPT-2` variants show the lowest effectiveness overall, but improves 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 improvements in NRP are 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`, 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 of `Llama-2`. `Llama-2 7B` 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 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 compared the preferences of our chosen ranking model with that of an expert annotator. Table 1 summarizes our results. For each of the 20 queries, the table indicates the rank assigned to LLM answers and document by (1) the expert and (2) `monoT5`. Based on these ranks, the table also shows the rank correlation for each query as measured by Kendall's τ. A consistent observation across almost all queries is that the expert assigned the best rank to `ChatGPT`, the second best to `Lamma-2 13B`, and the worst rank to `GPT-2`. For many queries, this is consistent with the ranks based on `monoT5` score. Most of the discrepancies between the expert and `monoT5` occurred with `GPT-2 XL` and the web document, as they often swapped places between the rankings. Across all queries, we observe a mean correlation of $\tau = 0.64$ between expert and monoT5 rank with a 95 % confidence interval of $[0.50, 0.78]$.

Query	Rank Positions of an Answer / Document					Correlation
qid	ChatGPT	Lla 13B	GPT-2 XL	GPT-2	Document	Kendall's τ
1						0.80
22						0.60
35						1.00
52						1.00
54						1.00
55						0.60
57						0.20
68						0.80
81						0.80
83						0.40
85						0.80
94						0.40
95						0.80
96						0.40
97						0.60
101						0.60
102						0.80
114						-0.20
116						0.60
117						0.80

Table 1. Correlation between the expert (▨) and `monoT5` (▨). Each cell shows the ranks 1-5 from left to right. Equal ranks by the expert and `monoT5` are illustrated with an overlap in color and pattern (▨). Each query is reported by its ID (qid).

5   Discussion & Limitations

By studying the effect of prompting strategy and model size on NRP, we find that (1) effectiveness increases with model size and (2) instruction-tuned models benefit more from elaborate prompting. However, we also find that NRP has difficulty distinguishing between highly effective models. Using the MultiMedQA prompt, larger models consistently generated answers that were ranked above all documents D_q. Whether this is an artifact of the dataset we used or whether this is also the case in other applications, needs to be investigated in future research.

In addition to the quality of the dataset, the efficacy of NRP is influenced by the underlying ranking model for capturing a quality dimension. As a consequence, the biases of a specific ranking model need to be considered when using it to evaluate LLMs with NRP for that dimension. For lexical models like TF-IDF and DPH, the biases are apparent: they use term-frequency to measure relevance and are biased towards LLM answers that contain many terms present in the query. For more sophisticated ranking models, the biases may be less apparent. In our experiments, we did not observe a consistent bias for `monoT5`.

In our user study, we correlated the rankings produced by `monoT5` with those made by a healthcare professional. The high correlation between the expert and `monoT5` suggest that rankers trained on human-written documents retain their effectiveness for LLM answers, despite the different structure and length of these answers. Naturally, this finding is limited to the specific dataset we used in our experiments and by the fact that we collected annotations from only one expert.

While our proposed method offers a new way to automatically score LLMs on open-ended questions, it depends on the availability of high-quality human annotations and ranking models that can capture them. This limits NRP to applications where these are available, and therefore makes it challenging to apply to low-resource languages and tasks where annotations are costly to obtain.

6 Conclusions

We developed a method to scalably and automatically evaluate generated answers to open-ended 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.

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 answers. The ranking functions can even be used to discriminate between answers from similar LLMs with different prompting strategies. 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 focused on evaluating generated answers for consumer health search questions on the CLEF 2021 eHealth dataset. 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 (2024), https://doi.org/10.48550/arxiv.2404.04044

[2] Berberette, E., Hutchins, J., Sadovnik, A.: Redefining "Hallucination" in LLMs: Towards a Psychology-Unformed Framework for Mitigating Misinformation. CoRR (2024), https://doi.org/10.48550/arXiv.2402.01769

[3] Duh, K.: Ranking vs. Regression in Machine Translation Evaluation. In: Proceedings of the Third Workshop on Statistical Machine Translation, pp. 191–194, StatMT '08, Association for Computational Linguistics, USA (2008), ISBN 9781932432091, https://doi.org/10.5555/1626394.1626425

[4] Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., Auli, M.: ELI5: Long Form Question Answering. In: ACL, pp. 3558–3567 (2019), https://doi.org/10.18653/v1/P19-1346

[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

[6] Gienapp, L., Scells, H., Deckers, N., Bevendorff, J., Wang, S., Kiesel, J., Syed, S., Fröbe, M., Zuccon, G., Stein, B., Hagen, M., Potthast, M.: Evaluating Generative Ad Hoc Information Retrieval. In: Yang, G., Wang, H., Han, S., Hauff, C., Zuccon, G., Zhang, Y. (eds.) 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024), pp. 1916–1929, ACM (Jul 2024), ISBN 9798400704314, https://doi.org/10.1145/3626772.3657849

[7] Goeuriot, L., Suominen, H., Pasi, G., Bassani, E., Brew-Sam, N., Gonzalez-Saez, 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.) CLEF 2021 Working Notes, CEUR Workshop Proceedings Vol. 2936., vol. 2936, Guglielmo Faggioli and Nicola Ferro and Alexis Joly and Maria Maistro and Florina Piroi, Bucharest, Romania (Sep 2021), URL https://hal.science/hal-04109627

[8] 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

[9] Khattab, O., Zaharia, M.: ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39–48, SIGIR '20, Association for Computing Machinery, New York, NY, USA (2020), ISBN 9781450380164, https://doi.org/10.1145/3397271.3401075

[10] 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

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

[12] Li, M., Jiang, A., Wang, M.: Listwise Approach to Learning to Rank for Automatic Evaluation of Machine Translation. In: Way, A., Sima'an, Khalil andForcada, M.L. (eds.) Proceedings of Machine Translation Summit XIV: Papers, Nice, France (Sep 2013)

[13] 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

[14] Nogueira, R., Jiang, Z., Pradeep, R., Lin, J.: Document Ranking with a Pretrained Sequence-to-Sequence Model. In: Cohn, T., He, Y., Liu, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 708–718, Association for Computational Linguistics, Online (Nov 2020), https://doi.org/10.18653/v1/2020.findings-emnlp.63

[15] Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of GPT-4 on Medical Challenge Problems (Apr 2023), https://doi.org/10.48550/arXiv.2303.13375

[16] 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. In: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2024), ISBN 9781713871088, https://doi.org/10.5555/3600270.3602281

[17] Pradeep, R., Nogueira, R.F., Lin, J.: The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models. CoRR (2021), https://doi.org/10.48550/arXiv.2101.05667

[18] 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

[19] 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

[20] Sakai, T.: SWAN: A Generic Framework for Auditing Textual Conversational Systems. CoRR (2023), https://doi.org/10.48550/arXiv.2305.08290

[21] Santhanam, K., Khattab, O., Saad-Falcon, J., Potts, C., Zaharia, M.: ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3715–3734, Association for Computational Linguistics, Seattle, United States (Jul 2022), https://doi.org/10.18653/v1/2022.naacl-main.272

[22] 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

[23] 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