Qeios

Peer Review

Review of: "MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems"

Ivan Sekulić[1]

1. Universita della Svizzera Italiana, Lugano, Switzerland

The work presented in the article is timely and important, as the rise of RAG-based methods continues. The benchmark consists of 110 human-generated, multi-turn interactions, with each conversational turn involving a passage retrieval and an answer generation step. I'll first talk about the bigger issues I see in this article, moving towards lesser issues and strengths of the paper. Please don't mistake my brevity for rudeness—I'm keeping my comments concise for clarity and efficiency. I appreciate the effort that went into this work and am happy to be involved in reviewing this article.

The main problem with the paper is that it lacks precise positioning within relevant prior work. First, the authors mention RAG as a *task* for LLMs, then as a type of multi-turn conversations. However, I'd argue RAG is a technique that can, but doesn't need to, involve LLMs. RAG, as a method or technique, can be used to tackle various challenges. However, these challenges need to be specified -- is the benchmark aimed towards tackling challenges in task-oriented dialog (e.g., in the abstract, "842" tasks -- this language is usually used when a dialog system aims to fulfil a user's goal, such as booking an appointment)? Or is it geared towards conversational information-seeking; if yes, is it more on the side of conversational search or conversational Q&A? You also mention "chat-based assistants", which is rather vague. See [1, 2, 3] for further info.

Strengths:

- Timely and important benchmark;

 doi.org/10.32388/MXLLH8

- High-quality human annotation efforts;
- Large number of conversations in the benchmark (110, with on average 7.7 turns each);
- The benchmark contains both human- and LLM-based evaluation;
- Diverse types of questions;
- A well-designed GitHub repository with clear instructions.

Limitations:

- Intro: you mention "challenging aspects" that a "multi-turn conversation benchmark" should cover, retrieval and generation. However, "The generator should struggle to answer many of the questions correctly" is unclear, and it seems to be aimed towards designing a benchmark just so certain methods fail on it, not the one that actually replicates a real-world task and is grounded in previous work.
- Intro: "Our benchmark was constructed using a novel process" -- how precisely is it novel? How is it different from, e.g., the way TREC-based benchmarks are constructed?
- Intro: "mtRAG is the first end-to-end human-generated multi-turn RAG benchmark" -- I'm not sure this is true, see below.
- Related work section needs to be significantly expanded; you can use the reference below to get you started.
- Related to overlooking prior work, the authors do mention iKAT as a benchmark not involving active retrieval and multi-domain components. However, iKAT has a collection of 100M+ passages that the system needs to query at each conversational turn in order to acquire a set of passages to generate the final answer from. Also, it is open-domain, so essentially it covers multiple domains (I'm curious about your thoughts here, though, as it may be understood in different ways).
- Moreover, CAsT [4, 5] is overlooked. It is a comprehensive benchmark involving both retrieval from a multi-million document collection and generation of a final response.
- Table 1 can also include the number of passages in each of the benchmarks.
- Sect. 3: Creation of the initial response by the LLM is not clear -- what is the input, and how is it repaired precisely?
- Sections 3 and 4 should be merged.
- 4: How was the seed question selected?
- 5.1: Add a citation for Elser. BM25 is also a sparse retriever; distinguish them better.

- 5.1 and results on retrievers: "Since we use Elser for retrieval during data creation, there may be some biases towards Elser" --> there certainly are, and these biases arguably deem your retriever comparisons irrelevant.
- 5.2: missing citation [6] for query rewriting.
- 5.3: retrieval results are heavily biased. Were the methods' parameters tuned?
- 5.3: Table 4: it's not clear why there is a distinction of "Turn 1" and "> Turn 1". Explain and draw conclusions or display all turns in a plot.
- 6.1: Section 5 is focused on retrieval, yet now we're talking about "Retrieval Settings". Improve paper organization.
- 6: How can you compare results from different retrieval settings when the "Reference+RAG" setting involves only a part of the dataset? Comparison to other settings can only happen on the exact subset of the 426 tasks in this setting. From what I understood, Table 5 contains results for "reference" and "RAG" on the full benchmark, while "Reference+RAG" is evaluated only on 426 samples, or am I mistaken?
- 6.2 should come before we talk about the results of the LLMs, as well as 6.3.
- 7: what was the inter-annotator agreement metric that was used?
- Merge 7 and 8 under "evaluation". I'd like to see better structuring on the comparison of the automatic evaluation and human evaluation.
- 9: see [7] for potential similarities between your approach to synthetic data creation and user simulation for conversational search systems.
- 10: Draw actual conclusions and takeaway messages; structure them. Too short.

Lesser limitations and notes:

- Abstract mentions "several additional challenges" -- such as?
- Abstract mentions "several real-world properties" -- such as?
- Intro: "important and popular field" -- I'd argue it's more of a topic than a field?
- Ensure consistent capitalization of terms (e.g., Large Language Models (LLMs) vs. Retrieval-augmented generation (RAG))—choose one format and apply it uniformly.

References:

1. Anand et al., 2020. [**Conversational Search**--A Report from **Dagstuhl** Seminar 19461] (https://arxiv.org/abs/2005.08658)

2. Zamani et al., 2023 [**Conversational information seeking**] (https://www.nowpublishers.com/article/Details/INR-081)

3. Dalton et al., 2022 [**Conversational information seeking**: theory and application] (https://dl.acm.org/doi/abs/10.1145/3477495.3532678)

4. Dalton et al., 2020 [**TREC CAsT** 2019: The conversational assistance track overview] (https://arxiv.org/abs/2003.13624)

5. Owoicho et al., 2022 [**TREC CAsT** 2022: Going Beyond User Ask and System Retrieve with Initiative and Response Generation.](https://trec.nist.gov/pubs/trec31/papers/Overview_cast.pdf)

6. Vakulenko et al., 2021 [A comparison of question **rewriting** methods for conversational passage retrieval](https://link.springer.com/chapter/10.1007/978-3-030-72240-1_43)

7. Owoicho et al., 2023 Exploiting simulated **user** feedback for conversational search: Ranking, rewriting, and beyond

Declarations

Potential competing interests: No potential competing interests to declare.