Challenges in RAG Evaluation for Text Classification in Evidence Synthesis

Sagar Uprety
University College London
United Kingdom
s.uprety@ucl.ac.uk

Ailbhe Finnerty
University College London
United Kingdom
a.finnerty@ucl.ac.uk

James Thomas
University College London
United Kingdom
james.thomas@ucl.ac.uk

Abstract

We developed a Retrieval Augmented Generation (RAG) system for automating systematic reviews and evidence synthesis in healthcare interventions, specifically focusing on smoking cessation behavioral interventions. This addresses a critical challenge in evidence-based medicine: the time-intensive process of manually reviewing and categorizing intervention characteristics from research papers. Our system approaches this as a classification task, determining whether specific modes of intervention delivery are present in each study through a sophisticated combination of document retrieval and large language model (LLM) reasoning. This abstract presents select challenges we encountered during the evaluation of our RAG system.

Keywords

LLM, RAG, Evaluation

ACM Reference Format:
Sagar Uprety, Ailbhe Finnerty, and James Thomas. 2025. Challenges in RAG Evaluation for Text Classification in Evidence Synthesis. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Background

Systematic reviews in healthcare research, particularly in behavioral interventions for smoking cessation, require extensive manual effort to analyze and categorize intervention characteristics from research papers [4]. We developed an approach using Retrieval Augmented Generation (RAG) to assist in this process, focusing on identifying modes of intervention delivery by formulating it as a binary classification task. For each research paper, the system must determine whether specific modes of delivery (e.g., face-to-face, digital, wearable, etc.) were used in the intervention. We carry out standard RAG experiments on a dataset of 40 papers. The various steps included PDF parsing, chunking, vector database indexing, and retrieval followed by prompting an LLM. We experimented with different settings for chunking, chunk sizes, top K, re-ranking, query expansion, hybrid vs dense retrieval, and various prompting techniques, including chain of thought prompting.

2 Challenges in Evaluation

For our evaluation of the RAG system [1], we evaluate both the retrieval component utilising standard IR metrics like MRR [3], Precision@k, Recall@k, NDCG@k [5], and also the generation/classification component using Precision, Recall, and F1 per attribute/class.

One of the biggest challenges was that the retrieval evaluation metrics did not correlate with the classification evaluation [6]. An example is reported in table 1 where F1-score for the classification task did not improve even with a significant improvement in retrieval metrics. We analysed the LLM outputs of the classification task using chain-of-thought prompts [7] and also asking the LLM to cite the part of the retrieved documents which relates to the output class. We found in some cases, the annotation of ground truth chunks was not foolproof, there were many more chunks which contained information regarding the class, which were not labelled as ground truth. This penalised the retriever [2], hence even when we saw poor retrieval results, the overall classification results were not affected. We further verified this by using non-ground truth chunks as retrieved context in the prompt and found that the classification performance does not drop much.

Another challenge was the presence of conflicting evidence within the same chunk. For example, there are two classes of interventions - face to face and distance. Sometimes information about both these was present in the same chunk, which led the LLM to classify only one of them (we need to further analyse whether position bias plays a role here). We used a smaller chunk size setting in order to mitigate this issue but that led to lower overall classification performance. One possible solution we intend to try is to club together some intervention attributes (e.g. face to face, distance) into a single prompt for a multi-label classification task.

Another type of challenge was the presence of irrelevant chunks in the top k retrieved chunks which were distracting the LLM towards an incorrect classification [2]. While precision-related metrics can be a helpful indicator of the retrieval quality in such cases, the challenge is that not all irrelevant chunks equally contribute to distracting the LLM. It's worth investigating when and how irrelevant chunks degrade the downstream performance.

We continue investigation of these and other challenges in RAG evaluation and experiment with different solutions.

Setting	Recall@1	Precision@1	Recall@k	NDCG@k	F1
Prompt_1	0.16	0.22	0.48	0.60	0.76
Prompt_2	0.30	0.42	0.65	0.63	0.77

Table 1: Performance metrics before and after changing the query text.

Unpublished working draft. Not for distribution.

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, Washington, DC, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

References

[1] Alaofi, M., Arabzadeh, N., Clarke, C. L., and Sanderson, M. Generative information retrieval evaluation. In *Information Access in the Era of Generative AI*. Springer, 2024, pp. 135–159.

[2] Alinejad, A., Kumar, K., and Vahdat, A. Evaluating the retrieval component in llm-based question answering systems. *arXiv preprint arXiv:2406.06458* (2024).

[3] Craswell, N. *Mean Reciprocal Rank*. Springer US, Boston, MA, 2009, pp. 1703–1703.

[4] Finnerty Mutlu, A., Howes, E., Veall, C., Thomas, J., O'Mara-Eves, A., West, R., Johnston, M., and Michie, S. Automated information extraction for behavioural interventions: evaluation and reflections on interdisciplinary ai development [version 1; peer review: 1 approved with reservations]. *Wellcome Open Research 9*, 493 (2024).

[5] Järvelin, K., and Kekäläinen, J. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS) 20*, 4 (2002), 422–446.

[6] Salemi, A., and Zamani, H. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 2024), SIGIR '24, Association for Computing Machinery, p. 2395–2400.

[7] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (Red Hook, NY, USA, 2022), NIPS '22, Curran Associates Inc.

Unpublished working draft.
Not for distribution.