When it comes to *metrics*, the use of LLMs as automatic evaluative judges signifies a burgeoning trend, promising versatility and depth in generative outputs with reasoning on a large scale compared to human evaluation. However, using “LLMs as a Judge” [75] for responses presents challenges in aligning with human judgment, establishing effective grading scales, and applying consistent evaluation across varied use cases. Determining correctness, clarity, and richness can differ between automated and human assessments. Moreover, the effectiveness of example-based scoring can vary, and there’s no universally applicable grading scale and prompting text, complicating the standardization of “LLM as a Judge”. [33]

In addition to the challenges mentioned above, it is important to consider the resource-intensive nature [76] of using Large Language Models (LLMs) for data generation and validation. RAG benchmarks must balance the need for thorough evaluation with the practical constraints of limited computational resources. As such, it is desirable to develop evaluation methodologies that can effectively assess RAG systems using smaller amounts of data while maintaining the validity and reliability of the results.

5 Conclusion

This survey systematically explores the complexities of evaluating RAG systems, highlighting the challenges in assessing their performance. Through the proposed *A Unified Evaluation Process of RAG*, we outline a structured approach to analyzing RAG evaluations, focusing on targets, datasets and measures. Our analysis emphasizes the need for targeted benchmarks that reflect the dynamic interplay between retrieval accuracy and generative quality and practical considerations for real-world applications. By identifying gaps in current methodologies and suggesting future research directions, we aim to contribute to more effective, and user-aligned benchmarks of RAG systems.

References

1. Balaguer, A., Benara, V., Cunha, R.L.d.F., Filho, R.d.M.E., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L.O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., Chandra, R.: RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. Tech. rep. (Jan 2024), http://arxiv.org/abs/2401.08406, arXiv:2401.08406 [cs] type: article
2. Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., Abdelrazek, M.: Seven failure points when engineering a retrieval augmented generation system (Jan 2024). https://doi.org/10.48550/ARXIV.2401.05856
3. Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., Nyczyk, P., Hoefler, T.: Graph of thoughts: Solving elaborate problems with large language models. Proceedings of the AAAI Conference on Artificial Intelligence 2024 (AAAI'24) (Aug 2023). https://doi.org/10.48550/ARXIV.2308.09687
4. Blagojevic, V.: Enhancing RAG Pipelines in Haystack: Introducing DiversityRanker and LostInTheMiddleRanker (Aug 2023), https://towardsdatascience.com/enhancing-rag-pipelines-in-haystack-45f14e2bc9f5
5. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al.: A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology **15**(3), 1–45 (2024)
6. Chen, J., Lin, H., Han, X., Sun, L.: Benchmarking large language models in retrieval-augmented generation (Sep 2023). https://doi.org/10.48550/ARXIV.2309.01431
7. Cuconasu, F., Trappolini, G., Siciliano, F., Filice, S., Campagnano, C., Maarek, Y., Tonellotto, N., Silvestri, F.: The power of noise: Redefining retrieval for rag systems (Jan 2024). https://doi.org/10.48550/ARXIV.2401.14887
8. Deng, Y., Zhang, W., Chen, Z., Gu, Q.: Rephrase and respond: Let large language models ask better questions for themselves (Nov 2023). https://doi.org/10.48550/ARXIV.2311.04205
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
10. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: Eraser: A benchmark to evaluate rationalized nlp models
11. Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of Wikipedia: Knowledge-powered conversational agents. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)
12. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024)
13. DuckDuckGo: DuckDuckGo — Privacy, simplified. (2024), https://duckduckgo.com//home
14. Es, S., James, J., Espinosa-Anke, L., Schockaert, S.: Ragas: Automated evaluation of retrieval augmented generation (Sep 2023). https://doi.org/10.48550/ARXIV.2309.15217