15. Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., Chen, D.: MRQA 2019 shared task: Evaluating generalization in reading comprehension. In: Fisch, A., Talmor, A., Jia, R., Seo, M., Choi, E., Chen, D. (eds.) Proceedings of the 2nd Workshop on Machine Reading for Question Answering. pp. 1–13. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-5801, https://aclanthology.org/D19-5801
16. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., Wang, H.: Retrieval-Augmented Generation for Large Language Models: A Survey. Tech. rep. (Jan 2024), http://arxiv.org/abs/2312.10997, arXiv:2312.10997 [cs] type: article
17. 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. Tech. rep. (Nov 2023), http://arxiv.org/abs/2311.04694, arXiv:2311.04694 [cs] type: article
18. Google: Programmable Search Engine | Google for Developers (2024), https://developers.google.com/custom-search
19. Gottschalk, S., Demidova, E.: Eventkg: A multilingual event-centric temporal knowledge graph (Apr 2018). https://doi.org/10.48550/ARXIV.1804.04526
20. Hofstätter, S., Chen, J., Raman, K., Zamani, H.: FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1437–1447. SIGIR '23, Association for Computing Machinery, New York, NY, USA (Jul 2023). https://doi.org/10.1145/3539618.3591687, https://doi.org/10.1145/3539618.3591687
21. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. Tech. rep. (Oct 2021). https://doi.org/10.48550/arXiv.2106.09685, http://arxiv.org/abs/2106.09685, arXiv:2106.09685 [cs] type: article
22. Huang, J., Shao, H., Chang, K.C.C., Xiong, J., Hwu, W.m.: Understanding jargon: Combining extraction and generation for definition modeling. In: Proceedings of EMNLP (2022)
23. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions (Nov 2023). https://doi.org/10.48550/ARXIV.2311.05232
24. Huang, Y., Huang, J.: A survey on retrieval-augmented text generation for large language models (Apr 2024). https://doi.org/10.48550/ARXIV.2404.10981
25. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
26. Kamalloo, E., Thakur, N., Lassance, C., Ma, X., Yang, J.H., Lin, J.: Resources for brewing beir: Reproducible reference models and an official leaderboard (2023)
27. Kasai, J., Sakaguchi, K., Takahashi, Y., Bras, R.L., Asai, A., Yu, X., Radev, D., Smith, N.A., Choi, Y., Inui, K.: Realtime qa: What's the answer right now? (Jul 2022). https://doi.org/10.48550/ARXIV.2207.13332, https://arxiv.org/abs/2207.13332
28. Khattab, O., Zaharia, M.: Colbert: Efficient and effective passage search via contextualized late interaction over bert (Apr 2020). https://doi.org/10.48550/ARXIV.2004.12832
29. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: A benchmark for question

answering research. Transactions of the Association for Computational Linguistics **7**, 453–466 (2019). https://doi.org/10.1162/tacl_a_00276, https://doi.org/10.1162/tacl_a_00276

30. Lahitani, A.R., Permanasari, A.E., Setiawan, N.A.: Cosine similarity to determine similarity measure: Study case in online essay assessment. In: 2016 4th International Conference on Cyber and IT Service Management. pp. 1–6 (2016). https://doi.org/10.1109/CITSM.2016.7577578
31. Lanchantin, J., Toshniwal, S., Weston, J., Szlam, A., Sukhbaatar, S.: Learning to reason and memorize with self-notes (May 2023). https://doi.org/10.48550/ARXIV.2305.00833
32. LangChain: Evaluating rag architectures on benchmark tasks (Nov 2023), https://langchain-ai.github.io/langchain-benchmarks/notebooks/retrieval/langchain_docs_qa.html
33. Leng, Q., Uhlenhuth, K., Polyzotis, A.: Best Practices for LLM Evaluation of RAG Applications (Dec 2023), https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG
34. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. pp. 9459–9474. NIPS'20, Curran Associates Inc., Red Hook, NY, USA (Dec 2020)
35. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Tech. rep. (Apr 2021), http://arxiv.org/abs/2005.11401, arXiv:2005.11401 [cs] type: article
36. Liang, X., Song, S., Niu, S., Li, Z., Xiong, F., Tang, B., Wy, Z., He, D., Cheng, P., Wang, Z., Deng, H.: Uhgeval: Benchmarking the hallucination of chinese large language models via unconstrained generation. arXiv preprint arXiv:2311.15296 (2023)
37. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
38. Liu, Y., Huang, L., Li, S., Chen, S., Zhou, H., Meng, F., Zhou, J., Sun, X.: Recall: A benchmark for llms robustness against external counterfactual knowledge (Nov 2023). https://doi.org/10.48550/ARXIV.2311.08147
39. Lyu, Y., Li, Z., Niu, S., Xiong, F., Tang, B., Wang, W., Wu, H., Liu, H., Xu, T., Chen, E., Luo, Y., Cheng, P., Deng, H., Wang, Z., Lu, Z.: Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models (Jan 2024). https://doi.org/10.48550/ARXIV.2401.17043
40. Microsoft: Web Search API | Microsoft Bing, https://www.microsoft.com/en-us/bing/apis/bing-web-search-api
41. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T.,