

Evaluating Retrieval-Augmented Generation Systems on Unanswerable, Uncheatable, Realistic, Multi-hop Queries

Gabrielle Kaili-May Liu¹, Bryan Li², Arman Cohan¹, William Gantt Walden^{3,4}, and Eugene Yang^{3,4}

¹ Yale University

² University of Pennsylvania

³ Human Language Technology Center of Excellence

⁴ Johns Hopkins University

kaili.liu@yale.edu, {wwalden1,eugene.yang}@jhu.edu

Abstract. Real-world use cases often present RAG systems with complex queries for which relevant information is missing from the corpus or is incomplete. In these settings, RAG systems must be able to reject unanswerable, out-of-scope queries and identify failures of retrieval and multi-hop reasoning. Despite this, existing RAG benchmarks rarely reflect realistic task complexity for multi-hop or out-of-scope questions, which often can be cheated via disconnected reasoning (i.e., solved without genuine multi-hop inference) or require only simple factual recall. This limits the ability for such benchmarks to uncover limitations of existing RAG systems. To address this gap, we present the first pipeline for automatic, difficulty-controlled creation of uncheatable, realistic, unanswerable, and multi-hop queries (CRUMQs), adaptable to any corpus and domain. We use our pipeline to create CRUMQs over two popular RAG datasets and demonstrate its effectiveness via benchmark experiments on leading retrieval-augmented LLMs. Results show that compared to prior RAG benchmarks, CRUMQs are highly challenging for RAG systems and achieve up to 81.0% reduction in cheatability scores. More broadly, our pipeline offers a simple way to enhance benchmark difficulty and realism and drive development of more capable RAG systems.

Keywords: multi-hop QA · unanswerability evaluation · synthetic data

1 Introduction

Retrieval Augmented Generation (RAG) [11, 23] is a powerful approach for many NLP tasks, enabling LLMs to respond to diverse user requests by leveraging an external document collection. While RAG is highly effective at increasing model credibility [15], mitigating hallucinations, and improving response quality [12], there remains a need to better understand how such systems handle complex, multi-part queries when available information from the corpus is insufficient. In particular, RAG systems must be able to appropriately reject unanswerable

queries and localize retrieval or reasoning failures when responding to multi-hop requests [29, 37]. These capabilities are crucial for reliable deployment of RAG systems in high-stakes domains where information is often missing or incomplete.

Existing RAG benchmarks [14, 44] rarely evaluate systems’ ability to handle *realistic* multi-hop or unanswerable queries. Multi-hop RAG benchmarks [21, 37] focus on synthetic task domains or suffer from disconnected reasoning [38, 39] wherein shortcuts can be exploited to achieve correct answers, while unanswerable RAG benchmarks [29] remain limited to simplistic factual recall settings.

To address these deficiencies, we present the first pipeline for automatic generation of uncheatable, realistic, unanswerable, multihop queries (CRUMQs), which are robust against reasoning shortcuts, target content beyond retrieval-augmented LLMs’ training data cutoff dates, and can be tailored to any document corpus. We leverage recent insights in synthetic data generation to ensure coverage of diverse task types and complexity levels, with benchmark difficulty easily controllable via the distribution of in- vs. out-of-scope hops per question.

We use our pipeline to create CRUMQs over two popular RAG datasets and showcase its efficacy through experiments on leading retrieval-augmented LLMs. Analysis reveals that CRUMQs pose notable difficulty even for RAG systems employing SOTA models such as GPT-5 [28]. We further show that CRUMQs are significantly less cheatable via disconnected reasoning than prior multi-hop RAG benchmarks, achieving up to an 81.0% decrease in cheatability. Overall, our work contributes to a better understanding of RAG systems’ limitations in handling unanswerable queries. Beyond driving the development of stronger and more capable RAG systems, our pipeline opens the door to automatically increasing the difficulty of existing datasets, addressing the challenge of benchmark longevity.

2 Related Work

A few studies have examined RAG system performance on queries which either require multi-hop reasoning or are beyond the scope of the relevant document collection [34]. However, these RAG benchmarks are restricted to singular domains and reflect low task complexity: multi-hop queries are fully answerable given the associated corpus and involve ≤ 4 hops, while the out-of-scope queries generally target short factual recall tasks. For instance, MultiHop-RAG [37] targets the news domain, with queries based on 2-4 document chunks, but nearly 90% of questions can be solved by GPT-4—reflecting low difficulty and reduced benchmark value. MHTS [21] generates difficulty-controllable multi-hop RAG queries, but QA pairs are created over a *single* novel to evaluate a *single* RAG system, limiting generalizability. On the other hand, UAEval4RAG [29] presents a framework to create out-of-database and inherently unanswerable RAG requests, yet resulting queries exhibit low complexity, lack difficulty modulation, and may overlap with models’ parametric knowledge due to limited data verification. Beyond these, common multi-hop benchmarks used for RAG tend to suffer from disconnected reasoning, not involve unanswerability, or reflect limited response formats (e.g., only short entities) [19, 32, 39–41]. Other QA datasets present



Fig. 1. Overview of the CRUMQs generation pipeline.

both answerable and unanswerable queries, but these adopt narrow task formulations, are not multi-hop, and/or do not involve retrieval [16, 30, 33, 35, 36, 43]. In contrast, we present the first pipeline for creating queries tailored to a given corpus that are *both* unanswerable *and* multi-hop—*and* of realistic complexity.

3 Method

We follow the pipeline shown in Fig. 1 to generate CRUMQs. We begin by extracting relevant topic keyphrases over the provided document collection. In this work, we assume the collection is a RAG corpus D with associated information-seeking requests and gold retrieved documents from D per request. However, our pipeline may be generalized by using synthetic requests or topic modeling.

Step I. Topics are extracted via two simple steps. A frontier LLM is few-shot prompted to extract short keyphrases from each information-seeking request.⁵ To obtain finer-grained topics, we then pair each initial topic with a gold document for the request and employ a second few-shot prompt to extract document-grounded topics. To ensure topics are sufficiently distinct and have good coverage of the corpus, we perform a deduplication step via embedding similarity. We use the `BAAI/bge-large-en-v1.5` embedder with a similarity threshold of 0.95.

Step II. To collect relevant information that is likely beyond the given corpus and training data cutoff date for retrieval-augmented LLMs, we next crawl the N_e most related recent articles for each topic from the external sources Google News [4], bioRxiv [2], chemRxiv [3], medRxiv [5], arXiv [1], and PubMed [7].

Step III. The externally sourced articles are then used to generate queries that are either *fully* unanswerable (relevant information is not found in D , only in the externally sourced documents) or *partially* unanswerable (relevant information is in D , but at least one key fact or detail required to provide a complete and correct answer is absent). To do so, we first split each gold article and each external article into 1,024-token chunks via LangChain [8], tracking for each chunk its source, URL, associated topic keyphrase, and associated request. Chunks are filtered for relevance to the topic and request via a binary LLM judgment. We then construct the set of all possible groups of 2–6 chunks such that at least one externally sourced chunk is in each group. As this set may be overly large in practice, we place a limit N_c on the number of contexts that are created for each total number of chunks and each ratio of external:gold chunks. To obtain seed queries, we prompt a strong generator LLM to systematically create up to 10 multi-hop

⁵ Our code and prompts will be released soon.

QA pairs using each multi-chunk context,⁶ making sure information is leveraged across chunk boundaries. For fully unanswerable queries we require all chunks in a context to be externally sourced; for partially unanswerable queries we require at least one gold chunk. We ensure diverse task formulations by adapting prompts from prior work in multi-document QA generation [25, 42].

Step IV. The unanswerability of each seed question is verified as in UAEval4RAG [29]: each question is used to retrieve the top 10 relevant chunks from the original corpus D , and LLM judgment is used to verify these chunks cannot answer the question. QA pairs that pass this verification are retained.

Step V. The data is finally filtered to ensure only truly multi-hop and high-quality queries remain. Following [24], we first annotate each QA pair with a synthetic chain-of-thought (CoT) explanation of the answer in the oracle setting (i.e., assuming both gold and externally sourced documents are available), and record the number of hops required to solve the question. Queries that do not adhere to the intended hop count are discarded.⁷ We then utilize a strong LLM to assess each unanswerable QA pair in the oracle setting according to the following criteria [13, 37]: context necessity, context sufficiency, answer correctness, answer uniqueness. Contextual criteria are additionally assessed assuming only the given corpus is available. We utilize the same scoring scale (Likert, 0-2) as [13]: only QA pairs which receive a score of at least 1 for all 6 criteria are retained.⁸

4 Experimental Setup

We demonstrate the efficacy of our pipeline in creating unanswerable, uncheatable, realistic, and multi-hop RAG queries by comparing against established RAG benchmarks. We consider UAEval4RAG [29] and MultiHop-RAG [37] as leading baselines, as they are the most recent works to perform comprehensive unanswerable and multi-hop RAG evaluation, respectively.

Datasets. Using our pipeline, we first generate CRUMQs over NeuCLIR [6] and TREC RAG 2025 [9], two popular RAG datasets which each consist of a large document collection with associated user requests. We focus on English texts, set $N_e = 200$ and $N_c = 50$, and use Llama3.3-70B-Instruct [18] as the generation and verification model to balance costs and quality.⁹ This leads to a total of 3,048 CRUMQs. We next use UAEval4RAG to generate baseline out-of-database RAG queries over NeuCLIR and TREC RAG 2025, leading to 7,559 out-of-database queries. Finally, we utilize the MultiHop-RAG dataset as-is, which consists of 2,556 multi-hop RAG queries and their associated gold contexts.

Unanswerability Evaluation. To verify the unanswerability value of our pipeline, we adopt the same experimental setup as in [29] to benchmark four leading RAG

⁶ We found that QA generation with multi-chunk contexts was more fruitful than with extracted claims [37] or entity-relation triplets [22] from document chunks.

⁷ We retain a selection of single-hop queries in §4 for comparability to prior works.

⁸ We manually review a subset of examples to validate the LLM verification results.

⁹ As we show in §5, creating effective CRUMQs does depend on proprietary models.

Table 1. Performance of RAG systems on CRUMQs vs. UAEval4RAG queries, using GPT-4o for generation. \dagger denotes Holm-Bonferroni corrected significance ($p < 0.05$).

Dataset	Embedding	Retrieval	Reranker	Rewriting	Accep. \uparrow	Unans. \uparrow	Clar. \uparrow	Acc. \uparrow
UAEval4RAG	Cohere	Vector	None	None	0.43	0.37	0.03	0.34
	Cohere	Vector	Cohere	HyDE	0.50	0.94	0.05	0.01
	BGE	Vector	Cohere	None	0.50	0.94	0.05	0.01
	OpenAI	Vector	Cohere	HyDE	0.42	0.45	0.05	0.28
CRUMQs	Cohere	Vector	None	None	0.34 \dagger	0.48 \dagger	0.06 \dagger	0.18 \dagger
	Cohere	Vector	Cohere	HyDE	0.29 \dagger	0.79 \dagger	0.21 \dagger	0.00 \dagger
	BGE	Vector	Cohere	None	0.29 \dagger	0.80 \dagger	0.20 \dagger	0.00 \dagger
	OpenAI	Vector	Cohere	HyDE	0.23 \dagger	0.66 \dagger	0.05 \dagger	0.06 \dagger

systems on the CRUMQs and UAEval4RAG queries, downsampled to 3,048 examples for fair comparison. We assume no access to external documents and use LlamaIndex [27] for RAG system implementation. We additionally evaluate the performance of RAG systems that use leading proprietary LLMs Gemini-2.5-Pro [17] and GPT-5 for generation, in addition to GPT-4o as in [29].

Cheatability Evaluation. To demonstrate the robustness of our queries against reasoning shortcuts, we compare our CRUMQs (downsampled to 2,556 examples) against MultiHop-RAG [37]. Following [38], we first obtain LLM predictions for each dataset in the oracle setting on the tasks of answer prediction and paragraph-level support identification. We use Llama3.1-8B-Instruct, Llama3.3-70B-Instruct, GPT-4o, and Gemini-2.5-Pro to ensure coverage of diverse model types and sizes. We then repeat the same experiment on the DiRe probe [38] of each dataset, which is a model-agnostic probe to gauge the extent of disconnected reasoning; additional details are in [38].

Metrics. For unanswerability evaluation, we adopt the metrics of acceptable ratio, unanswered ratio, and ask-for-clarification ratio from [29]. We additionally score accuracy by running LLM judgments of semantic equivalence between target and predicted answers (Gemini-2.0-Flash prompted as in [25]). For cheatability evaluation, we compute the average F1 score for each model \times dataset \times task setting as in [38, 39]. The cheatability of each dataset is then measured as the ratio of F1 scores in the probe vs. non-probe settings, which represents the percentage of model performance attributable to disconnected reasoning.

5 Results and Analysis

We report unanswerability evaluation results in Tables 1 and 2. Consistent with the findings in [29], we observe in Table 1 that no single system achieves the best performance on both CRUMQs and UAEval4RAG queries, with different system configurations able to yield similar results. Importantly, *CRUMQs are harder than UAEval4RAG queries*. All systems respond less acceptably to CRUMQs than UAEval4RAG queries. Moreover, CRUMQs pose notable difficulty for leading proprietary LLM-based RAG systems, which provide much fewer acceptable responses for CRUMQs versus UAEval4RAG queries (Table 2). While queries

Table 2. Performance of proprietary model RAG systems on CRUMQs vs. UAEval4RAG queries. As in [29], no reranking or rewriting is used. \dagger denotes Holm–Bonferroni corrected significance ($p < 0.05$).

Dataset	LLM	Embedding	Retrieval	Accep. \uparrow	Unans. \uparrow	Clar. \uparrow	Acc. \uparrow
UAEval4RAG	Gemini-2.5-Pro	OpenAI	Ensemble	0.48	0.51	0.00	0.39
	GPT-5	OpenAI	Ensemble	0.28	0.31	0.07	0.33
	GPT-4o	OpenAI	Ensemble	0.67	0.01	0.43	0.26
CRUMQs	Gemini-2.5-Pro	OpenAI	Ensemble	0.23 \dagger	0.23 \dagger	0.03 \dagger	0.19 \dagger
	GPT-5	OpenAI	Ensemble	0.28\dagger	0.03 \dagger	0.35 \dagger	0.18 \dagger
	GPT-4o	OpenAI	Ensemble	0.24 \dagger	0.28 \dagger	0.01 \dagger	0.12 \dagger

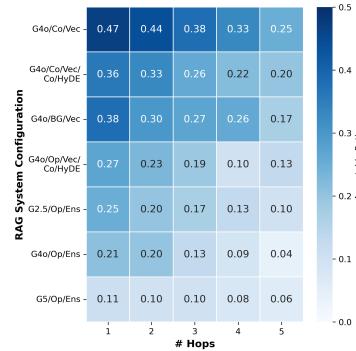


Fig. 2. Acceptable ratios of RAG systems on CRUMQs across hop counts. Performance drops with more hops.

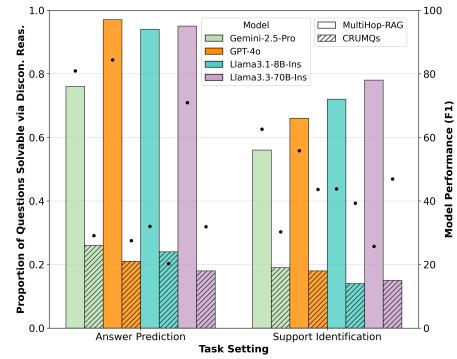


Fig. 3. DiRe F1 score ratios (\downarrow) across benchmarks and tasks. Black points denote accuracy per model per task (values on right axis).

from both systems occasionally capture parametric knowledge (i.e., answerable queries are produced, indicated by nonzero accuracy), this proportion is significantly lower for CRUMQs than for UAEval4RAG queries. Finally, CRUMQs with greater hop counts lead to consistently lower acceptable ratios across systems (Fig. 2), indicating easily controllable difficulty level for queries. Overall, our pipeline enables creation of harder out-of-database queries with controllable difficulty, which can be used to more effectively differentiate RAG systems.

Compared to prior multi-hop RAG benchmarks, CRUMQs are significantly more difficult and much less cheatable, with all comparisons significant after Holm–Bonferroni correction ($p < 0.05$). As indicated by the black points in Fig. 3, we observe that CRUMQs pose a much greater challenge for leading LLMs than MultiHop-RAG across all tasks, with models achieving answer prediction scores up to only 31.9 versus up to 84.4 for MultiHop-RAG. Moreover, up to 96.7% of queries in MultiHop-RAG can be answered via disconnected reasoning, compared to only up to 23.6% of our CRUMQs. A similar trend is observed for the support identification task. These results confirm the efficacy of our pipeline for creation of challenging multi-hop RAG queries.

6 Conclusion and Future Work

In this work, we introduced an adaptable framework that is the first of its kind for creating highly challenging, difficulty-controllable, unanswerable and multi-hop RAG queries that cannot be easily cheated. Experiments demonstrate that our pipeline effectively addresses deficiencies of existing RAG benchmarks. In addition to advancing the development of stronger RAG systems, our work paves the way toward automatically increasing benchmark difficulty through data augmentation. Future work may explore the performance of multi-hop-oriented RAG systems [10, 20, 26, 31] on CRUMQs, along with analysis of systems’ ability to localize sources of unanswerability and signal uncertainty in such settings.

Acknowledgments. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2139841. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. arxiv. <https://www.arxiv.org/>
2. biorxiv. <https://biorxiv.org>
3. chemrxiv. <https://www.chemrxiv.org/>
4. Google news. <https://news.google.com/>
5. medrxiv. <https://www.medrxiv.org/>
6. Neuclir corpus. <https://ir-datasets.com/neuclir.html>
7. Pubmed. <https://pubmed.ncbi.nlm.nih.gov/>
8. Tokentextsplitter, https://python.langchain.com/api_reference/text_splitters/base/langchain_text_splitters.base.TokenTextSplitter.html#tokentextsplitter
9. Trec 2025 rag corpus. <https://trec-rag.github.io/annoucements/2025-rag25-corpus/>
10. Agrawal, R., Asrani, M., Youssef, H., Narayan, A.: Scmrag: Self-corrective multihop retrieval augmented generation system for llm agents. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems. p. 50–58. AAMAS ’25, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2025)
11. Asai, A., Min, S., Zhong, Z., Chen, D.: Retrieval-based language models and applications. In: Chen, Y.N.V., Margot, M., Reddy, S. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts). pp. 41–46. Association for Computational Linguistics, Toronto, Canada (Jul 2023). <https://doi.org/10.18653/v1/2023.acl-tutorials.6>, <https://aclanthology.org/2023.acl-tutorials.6/>

12. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Milligan, K., Van Den Driessche, G.B., Lespiau, J.B., Damoc, B., Clark, A., De Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J., Elsen, E., Sifre, L.: Improving language models by retrieving from trillions of tokens. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 2206–2240. PMLR (17–23 Jul 2022), <https://proceedings.mlr.press/v162/borgeaud22a.html>
13. Chernogorskii, F., Averkiev, S., Kudraleeva, L., Martirosian, Z., Tikhonova, M., Malykh, V., Fenogenova, A.: Dragon: Dynamic rag benchmark on news (2025), <https://arxiv.org/abs/2507.05713>
14. Friel, R., Belyi, M., Sanyal, A.: Ragbench: Explainable benchmark for retrieval-augmented generation systems (2025), <https://arxiv.org/abs/2407.11005>
15. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling large language models to generate text with citations. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6465–6488. Association for Computational Linguistics, Singapore (Dec 2023). <https://doi.org/10.18653/v1/2023.emnlp-main.398>, <https://aclanthology.org/2023.emnlp-main.398/>
16. González Torres, J.J., Bîndilă, M.B., Hofstee, S., Szondy, D., Nguyen, Q.H., Wang, S., Englebienne, G.: Automated question-answer generation for evaluating RAG-based chatbots. In: Demner-Fushman, D., Ananiadou, S., Thompson, P., Ondov, B. (eds.) Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024. pp. 204–214. ELRA and ICCL, Torino, Italia (May 2024), <https://aclanthology.org/2024.cl4health-1.25/>
17. Google: Gemini 2.5 pro model card. <https://modelcards.withgoogle.com/assets/documents/gemini-2.5-pro.pdf> (2025)
18. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Diman, E., Smith, E.M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J.,

Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenbende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X.E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenber, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B.D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J.,

- Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N.P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., Ma, Z.: The llama 3 herd of models (2024), <https://arxiv.org/abs/2407.21783>
19. Ho, X., Duong Nguyen, A.K., Sugawara, S., Aizawa, A.: Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 6609–6625. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). <https://doi.org/10.18653/v1/2020.coling-main.580>, [https://aclanthology.org/2020.coling-main.580/](https://aclanthology.org/2020.coling-main.580)
 20. Hu, Y., Lei, Z., Dai, Z., Zhang, A., Angirekula, A., Zhang, Z., Zhao, L.: Cg-rag: Research question answering by citation graph retrieval-augmented llms (2025), <https://arxiv.org/abs/2501.15067>
 21. Lee, J., Kwon, D., Jin, K., Jeong, J., Sim, M., Kim, M.: Mhts: Multi-hop tree structure framework for generating difficulty-controllable qa datasets for rag evaluation (2025), <https://arxiv.org/abs/2504.08756>
 22. Lei, D., Li, Y., Li, S., Hu, M., Xu, R., Archer, K., Wang, M., Ching, E., Deng, A.: FactCG: Enhancing fact checkers with graph-based multi-hop data. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference

- of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 5002–5020. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). <https://doi.org/10.18653/v1/2025.nacl-long.258>, <https://aclanthology.org/2025.nacl-long.258/>
23. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* **33**, 9459–9474 (2020)
 24. Li, Y., Liang, S., Lyu, M., Wang, L.: Making long-context language models better multi-hop reasoners. In: Ku, L.W., Martins, A., Sriku-mar, V. (eds.) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 2462–2475. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). <https://doi.org/10.18653/v1/2024.acl-long.135>, <https://aclanthology.org/2024.acl-long.135/>
 25. Liu, G.K.M., Shi, B., Caciularu, A., Szpektor, I., Cohan, A.: MDCure: A scalable pipeline for multi-document instruction-following. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 29258–29296. Association for Computational Linguistics, Vienna, Austria (Jul 2025). <https://doi.org/10.18653/v1/2025.acl-long.1418>, <https://aclanthology.org/2025.acl-long.1418/>
 26. Liu, H., Wang, Z., Chen, X., Li, Z., Xiong, F., Yu, Q., Zhang, W.: Hoprag: Multi-hop reasoning for logic-aware retrieval-augmented generation (2025), <https://arxiv.org/abs/2502.12442>
 27. Liu, J.: LlamaIndex (11 2022). <https://doi.org/10.5281/zenodo.1234>, https://github.com/jerryjliu/llama_index
 28. OpenAI: Gpt-5 system card. <https://cdn.openai.com/gpt-5-system-card.pdf> (2025)
 29. Peng, X., Choubey, P.K., Xiong, C., Wu, C.S.: Unanswerability evaluation for retrieval augmented generation. *arXiv preprint arXiv:2412.12300* (2024)
 30. Peng, Z., Nian, J., Evfimievski, A., Fang, Y.: Eloq: Resources for enhancing llm detection of out-of-scope questions (2025), <https://arxiv.org/abs/2410.14567>
 31. Poliakov, M., Shvai, N.: Multi-Meta-RAG: Improving RAG for Multi-hop Queries Using Database Filtering with LLM-Extracted Metadata, p. 334–342. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-031-81372-6_25, http://dx.doi.org/10.1007/978-3-031-81372-6_25
 32. Rajpurkar, P., Jia, R., Liang, P.: Know what you don't know: Unanswerable questions for SQuAD. In: Gurevych, I., Miyao, Y. (eds.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. pp. 784–789. Association for Computational Linguistics, Melbourne, Australia (Jul 2018). <https://doi.org/10.18653/v1/P18-2124>, <https://aclanthology.org/P18-2124/>

33. Rosenthal, S., Sil, A., Florian, R., Roukos, S.: Clapnq: Cohesive long-form answers from passages in natural questions for rag systems (2024)
34. Shen, H., Yan, H., Xing, Z., Liu, M., Li, Y., Chen, Z., Wang, Y., Wang, J., Ma, Y.: Ragsynth: Synthetic data for robust and faithful rag component optimization (2025), <https://arxiv.org/abs/2505.10989>
35. Sun, Y., Yin, Z., Guo, Q., Wu, J., Qiu, X., Zhao, H.: Benchmarking hallucination in large language models based on unanswerable math word problem. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 2178–2188. ELRA and ICCL, Torino, Italia (May 2024), <https://aclanthology.org/2024.lrec-main.196/>
36. Tan, C., Shao, W., Xiong, H., Zhu, T., Liu, Z., Shi, K., Chen, W.: Uaqfact: Evaluating factual knowledge utilization of llms on unanswerable questions (2025), <https://arxiv.org/abs/2505.23461>
37. Tang, Y., Yang, Y.: Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries (2024)
38. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: Is multihop QA in DiRe condition? measuring and reducing disconnected reasoning. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 8846–8863. Association for Computational Linguistics, Online (Nov 2020). <https://doi.org/10.18653/v1/2020.emnlp-main.712>, [https://aclanthology.org/2020.emnlp-main.712/](https://aclanthology.org/2020.emnlp-main.712)
39. Trivedi, H., Balasubramanian, N., Khot, T., Sabharwal, A.: MuSiQue: Multihop questions via single-hop question composition. Transactions of the Association for Computational Linguistics **10**, 539–554 (2022). <https://doi.org/10.1162/tacl.a.00475>, <https://aclanthology.org/2022.tacl-1.31/>
40. Yang, X., Sun, K., Xin, H., Sun, Y., Bhalla, N., Chen, X., Choudhary, S., Gui, R.D., Jiang, Z.W., Jiang, Z., Kong, L., Moran, B., Wang, J., Xu, Y.E., Yan, A., Yang, C., Yuan, E., Zha, H., Tang, N., Chen, L., Scheffer, N., Liu, Y., Shah, N., Wanga, R., Kumar, A., tau Yih, W., Dong, X.L.: Crag – comprehensive rag benchmark. arXiv preprint arXiv:2406.04744 (2024), <https://arxiv.org/abs/2406.04744>
41. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D.: Hotpotqa: A dataset for diverse, explainable multi-hop question answering (2018), <https://arxiv.org/abs/1809.09600>
42. Zhang, J., Bai, Y., Lv, X., Gu, W., Liu, D., Zou, M., Cao, S., Hou, L., Dong, Y., Feng, L., Li, J.: LongCite: Enabling LLMs to generate fine-grained citations in long-context QA. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 5098–5122. Association for Computational Linguistics, Vienna, Austria (Jul 2025). <https://doi.org/10.18653/v1/2025.findings-acl.264>, <https://aclanthology.org/2025.findings-acl.264/>

43. Zhang, Q.W., Li, F., Wang, J., Qiao, L., Yu, Y., Yin, D., Sun, X.: Factguard: Leveraging multi-agent systems to generate answerable and unanswerable questions for enhanced long-context llm extraction (2025), <https://arxiv.org/abs/2504.05607>
44. Zhu, K., Luo, Y., Xu, D., Yan, Y., Liu, Z., Yu, S., Wang, R., Wang, S., Li, Y., Zhang, N., Han, X., Liu, Z., Sun, M.: RAGEval: Scenario specific RAG evaluation dataset generation framework. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8520–8544. Association for Computational Linguistics, Vienna, Austria (Jul 2025). <https://doi.org/10.18653/v1/2025.acl-long.418>, <https://aclanthology.org/2025.acl-long.418/>