Modeling Contextual Passage Utility for Multihop Question Answering

Akriti Jain, Aparna Garimella
Adobe Research, India
{akritij, garimell}@adobe.com

Abstract

Multihop Question Answering (QA) requires systems to identify and synthesize information from multiple text passages. While most prior retrieval methods assist in identifying relevant passages for QA, further assessing the utility of the passages can help in removing redundant ones, which may otherwise add to noise and inaccuracies in the generated answers. Existing utility prediction approaches model passage utility independently, overlooking a critical aspect of multi-hop reasoning, that the utility of a passage can be context-dependent, influenced by its relation to other passages—whether it provides complementary information, or forms a crucial link in conjunction with others. In this paper, we propose a light-weight approach to model contextual passage utility, accounting for inter-passage dependencies. We fine-tune a small transformer-based model to predict passage utility scores for multihop QA. We leverage the reasoning traces from an advanced reasoning model to capture the order in which passages are used to answer a question, to obtain synthetic training data. Through comprehensive experiments, we demonstrate that our utility-based scoring of retrieved passages leads to better reranking and downstream task performance compared to relevance-based reranking methods.

1 Introduction

Effective multihop question answering (QA) hinges on identifying not only relevant passages, but also those that are truly useful to answering the given question. While relevance merely signifies a topical connection between context and the query, utility reflects a passage's actual contribution to constructing the answer (Xu et al., 2025). Classic relevance labels often fail to predict QA success, and external relevance judgments tend to correlate poorly with QA performance (Salemi and Zamani, 2024). This insufficiency arises partly because retrieved passages may be individually relevant, yet only a subset of them are actually used within the specific inferential path required to reach the correct answer (Zhang et al., 2024). Consequently, utility judgments, rather than relevance alone, offer more valuable guidance for identifying ground-truth evidence and enhancing answer generation.

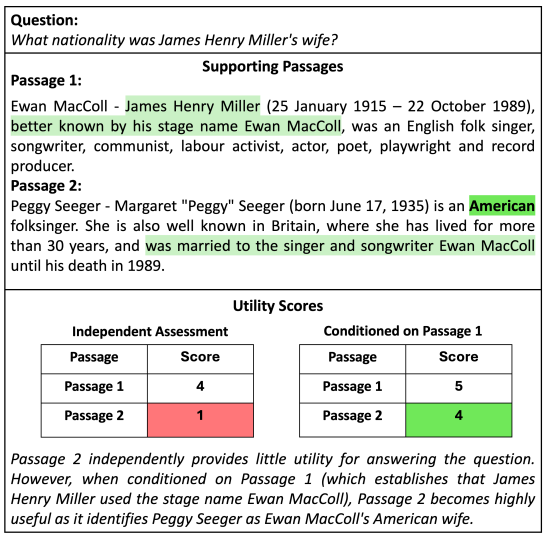

Figure 1: A multihop question from HotpotQA dataset (Yang et al., 2018): Passage 2 if considered independently does not seem useful to answer the question. However, conditioned on Passage 1, it becomes useful.

The work on passage utility modeling has been nascent. Perez-Beltrachini and Lapata (2025) proposed to incorporate factors such as QA model's accuracy and entailment scores between the passages and the answer, to predict the utility of a given passage. Despite the growing recognition, existing methods fall short in multihop QA. A main limitation is the tendency to assess passages in isolation, assuming their contribution is independent of other contextual information. This is particularly problematic in multihop scenarios where a passage's utility is frequently conditional—it may only become useful after another passage has established necessary context, such as a supporting entity in

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 464–471

one passage activating the usefulness of subsequent information (Figure 1). Thus, compositionality and order are paramount: a passage's utility is to be determined by its position within a reasoning chain and its dependencies on preceding content.

Our work aims to address this gap by proposing a novel approach to model passage utility by accounting for inter-passage dependencies. We synthetically generate utility ratings for passages, using the reasoning traces from an advanced reasoning model (o1) to capture the order in which the passages are used in addressing a question, which are then used for point-wise passage utility ratings on a scale of 1-5 annotated by GPT-4o. A RoBERTa-based model (Liu et al., 2019) is trained on these trace-level annotations to explicitly learn these contextual dependencies to predict utility scores.

Our contributions are threefold: **(1)** We address utility-aware passage ranking in multihop QA, which remains an underexplored topic in retrieval-augmented QA efforts. **(2)** We propose a light-weight scoring model that learns to predict utility, sensitive to inter-passage dependencies. We obtain path-dependent utility in multi-hop QA using ordered reasoning traces and LLM-generated ordinal scores. **(3)** Through experiments on various datasets, we illustrate that our approach significantly improves the identification of useful passage sets and enhances downstream QA performance compared to baselines that do not account for utility, or do so in a context independent manner.

2 Utility-Aware Contextual Passage Ranking

We formulate passage utility scoring in multihop question answering as a regression task. We maintain that the true utility of a passage p_i is inherently context-dependent, influenced by previously encountered passages $P_{<i}$ and the question q. Our objective is to predict a utility score $U(p_i|P_{<i}, q)$ on a scale of 1 to 5. A high score signifies a greater contribution of p_i towards answering q, given the contextual information derived from $P_{<i}$. Here, i indexes the position of p_i in an ordered set of retrieved passages.

Our passage utility predictor uses RoBERTa-large (Liu et al., 2019) as a regression model. It takes as input the question q and the current passage p_i being evaluated, and predicts a utility score indicating how useful p_i is for answering q. Our pipeline comprises of two stages: (1) synthetic data generation with utility scores for a given passage, conditioned on the other context, for a given question; and (2) utility predictor training using the RoBERTa-large model.

2.1 Training Data Generation

Our training data consists of (question, passage, utility score) triplets, where the utility scores are derived through a two-staged process designed to capture contextual dependencies. With the growing use of LLMs for generating high-quality synthetic training data (Wagner et al., 2025; Li et al., 2023; Ba et al., 2024), we adopt a similar strategy, allowing the reasoning capabilities of powerful LLMs to be distilled into smaller, more efficient models for downstream tasks.

Reasoning Trace Generation. To capture the compositional nature of information in multi-hop reasoning, we first generate explicit reasoning traces using a pre-trained reasoning model (OpenAI et al., 2024). Each passage within the context for a given question is tagged with a unique identifier (e.g., [Passage A], [Passage B]). The reasoning model is then prompted to produce a detailed reasoning trace that explicitly cites which passage(s) support each inferential step required to answer the question. These reasoning traces clarify how information is progressively integrated across multiple passages and illustrate the specific utility each passage contributes toward constructing the final answer in a multi-hop setting.

Utility Score Annotation. We employ another LLM, GPT-4o, to assign utility scores (ranging from 1 to 5) to each passage, conditioned on its function within the complete reasoning trace generated in the previous step. For each passage p_i, the scoring LLM is provided with the question q, the target passage p_i, and the full reasoning trace T associated with q. The LLM is prompted to evaluate p_i's utility by analyzing its usage in T. Specifically, it assesses: (1) whether p_i was explicitly cited as evidence in T; (2) the specific role p_i played according to T (e.g., providing initial facts, bridging information, or supplying final evidence); and (3) the criticality of this contribution to answering q as per T. The utility is then quantified on a 5-point scale, where 1 indicates a passage not being used to address q, and 5 represents an essential passage providing critical evidence without which the answer could not be derived according to the trace. This approach is to ensure that the utility scores U_i re-

flect not just the intrinsic relevance of a passage to the question, but its actual contribution within the specific reasoning context. These context-aware utility scores serve as supervision signals for training our regression model and their reliability is supported by a high upper-bound performance when used directly for passage ranking. We manually verify 200 examples across our different datasets; we note alignment between LLM-generated and human-rated passage utilities in over 95% cases, confirming their reliability as supervision signals (further details on both in Appendix A.1).

2.2 Training Procedure

We fine-tune the RoBERTa-large model by adding a sequence classification head configured for regression, which outputs a single scalar value representing the predicted utility score. While the RoBERTa model itself only receives the local (q, p_i) pair as direct input, it is trained to predict the context-dependent utility scores U_i (derived from the full reasoning trace), by minimizing the mean squared error (MSE) between its predicted utility scores $f_{RoBERTa}(p_i, q)$ and the LLM-annotated, context-dependent scores U_i.

$$L = \frac{1}{N}\sum_{j=1}^{N}(f_{RoBERTa}(p_{i,j}, q_j) - U_{i,j})^2 \quad (1)$$

where N is the total number of training examples. Here, j indexes individual training examples, each consisting of a question q_j, a passage $p_{i,j}$, and a utility score $U_{i,j}$.

3 Experimental Setup

We fine-tune the RoBERTa model for 3 epochs using the Adam optimizer (Kingma and Ba, 2017) with a weight decay of 0.01, along with a linear learning rate scheduler and a warmup ratio of 0.1. Training is performed with a per-device batch size of 8 and gradient accumulation steps set to 2. Mixed-precision (FP16) training is enabled. All experiments are conducted on a single NVIDIA A100 GPU (40GB).

3.1 Datasets

We evaluate our contextual passage utility scorer on three multihop QA benchmarks.

HotpotQA (Yang et al., 2018) features questions requiring 2-hop reasoning. Each question is typically accompanied by 2 gold supporting passages and 8 distractor passages.

MusiQue (Trivedi et al., 2022) is designed for compositional reasoning. The questions involve 2 to 4 reasoning hops. Each question includes 20 passages in total.

2WikiMultiHopQA (Ho et al., 2020) contains 2 or 4-hop questions constructed from Wikipedia, requiring the model to combine information from two different Wikipedia entities or facts to arrive at the answer. The total number of passages are 10.

The datasets we evaluate on already provide a fixed set of supporting and distractor passages for each question. This setup can be viewed as the output of a first-stage retriever, and hence, the number of candidate passages that our model scores is determined by each dataset's official construction.

We randomly select 5K question-passage pairs $(q; p_1, p_2, \ldots)$ for training, each annotated with a utility score from 1 to 5, based on the above setup.

3.2 Evaluation Metrics

Our evaluation targets three key aspects: identifying the correct set of ground-truth passages (P_G), ranking them effectively, and their utility in answering the question using the model-selected passages (P_M). For rank-based metrics, we consider the top 5 selected passages ($k = 5$). To assess how well P_M covers P_G, we use Precision@k (P@k), the fraction of selected passages that are correct; Recall@k (R@k), the fraction of ground-truth passages retrieved; and F1-Score@k (F1@k), the harmonic mean of P@k and R@k. Next, to evaluate the ranking quality of ground-truth passages within the top 5, we employ Normalized Discounted Cumulative Gain (NDCG@1, NDCG@5) (Järvelin and Kekäläinen, 2017), which measures overall ranking quality at the top 1 and top 5 positions. For downstream task performance, we assess if the top-k passages selected by our model are sufficient to answer the question. We report Exact Match (EM), which measures the percentage of predicted answers (generated using P_M) that exactly match a gold answer.

3.3 Baselines

We compare our model against several retrieval and reranking approaches. All rerankers operate on the same set of candidate passages per question. 1) **BM25:** A classical sparse retriever (Robertson and Zaragoza, 2009) that uses term-based scoring to rank candidate passages. 2) **Contriever:** An unsupervised dense retriever (Izacard et al., 2021). We

Dataset	Method	Multi-hop	P@2	R@2	F1@2	R@5	NDCG@1	NDCG@5	EM
HotpotQA	BM25*	✗	52.32	52.32	52.32	73.22	70.27	56.39	54.41
	Contriever*	✗	47.31	47.31	47.31	74.38	59.03	64.89	49.80
	MonoT5	✗	40.86	40.86	40.86	72.34	40.99	40.89	48.56
	MDR*	✓	62.04	62.04	62.04	85.62	80.05	79.24	73.57
	PromptRank	✓	50.79	50.79	50.79	71.01	70.28	66.31	61.05
	LLM	✓	70.20	70.20	70.20	86.10	85.22	73.49	84.86
	Ours	✓	**88.09**	**88.09**	**88.09**	**98.33**	**94.61**	**89.57**	**87.12**
MuSiQue	BM25*	✗	34.05	25.70	29.29	44.06	44.19	36.34	39.47
	Contriever*	✗	31.53	23.80	27.12	42.66	38.39	38.83	32.89
	MonoT5	✗	51.70	39.02	44.47	63.11	60.74	53.74	55.11
	MDR*	✓	51.11	41.19	45.62	66.65	63.01	61.75	61.30
	PromptRank	✓	42.20	32.90	36.99	48.84	57.84	48.23	32.77
	LLM	✓	45.51	37.04	40.17	50.06	64.36	50.06	**68.97**
	Ours	✓	**62.37**	**47.08**	**53.66**	**73.23**	**78.94**	**66.12**	68.51
2WikiMultiHopQA	BM25*	✗	49.75	40.82	44.85	67.68	61.84	52.49	32.26
	Contriever*	✗	26.86	22.04	24.21	41.48	33.97	36.21	18.38
	MonoT5	✗	61.70	50.63	55.62	80.79	65.60	62.59	50.85
	MDR*	✓	70.45	61.29	65.55	66.65	63.01	61.75	70.26
	PromptRank	✓	48.35	43.05	45.54	62.20	65.31	59.32	40.55
	LLM	✓	62.50	54.51	57.18	75.10	75.93	65.57	61.75
	Ours	✓	**94.51**	**83.66**	**88.75**	**99.02**	**97.70**	**95.23**	**79.44**

Table 1: Evaluation of contextual passage utility scoring on multi-hop QA datasets. EM = Exact Match; P@2 = Precision@2; R@2 = Recall@2; R@5 = Recall@5. Green highlights best performance; Red highlights second-best. * indicates retriever-based methods.

use its raw similarity scores for ranking, representing a strong dense retrieval baseline. 3) **MonoT5:** A T5-based neural reranker (Nogueira et al., 2020). We use a pre-trained model (castorini/monot5-base-msmarco) for pointwise passage relevance scoring. 4) **LLM-based Reranker (Zero-Shot):** GPT-4o prompted in a zero-shot manner to score the relevance of each candidate passage to the question. 3) Cross-Attention Reranking is a post-retrieval step integrated with the **Multi-Hop Dense Retrieval (MDR)** (Xiong et al., 2021) system. It takes the top-k passage sequences retrieved by MDR, prepends the original question to each, and uses a pre-trained Transformer encoder (like ELECTRA-large) to predict relevance scores. 5) **PROMPTRANK** (Kong et al., 2023) is an LLM-based reranker designed for few-shot multi-hop QA. It scores candidate document paths (sequences of documents) by measuring the large language model's conditional likelihood of generating the question given a prompt constructed from that path.

4 Results & Discussion

Table 1 shows the results across the three datasets. Our approach consistently outperforms all the baselines in both identifying the full set of useful passages (coverage - high R@k) and correctly ordering them (ranking - high NDCG scores). We observe significant improvements over even strong multi-hop–aware rerankers like MDR and PromptRank. On HotpotQA, for instance, our model achieves R@5 of 98.33, substantially higher than the strongest baseline (LLM scorer at 86.10), and NDCG@5 of 89.57 compared to the MDR Reranker 79.24. Similar gains are observed on MuSiQue and 2WikiMultiHopQA datasets as well.

Auxiliary Experiment. To further validate our approach, we conduct an auxiliary experiment with two key findings (detailed in Appendix A.2). In this setup, we fine-tune decoder-only models (LLaMA 3.2 1B and LLaMA 3.1 8B) to predict passage utility scores in two settings: (i) listwise scoring, where all candidate passages for a question are provided jointly, and (ii) pointwise scoring, where each passage is scored independently. Providing the full joint context in the listwise setup yields only marginal improvements over pointwise scoring. In contrast, our RoBERTa-based model, despite scoring one passage at a time, consistently outperforms these larger LLaMA models even when they have access to all candidate passages simultaneously. This indicates that the essential cross-passage dependencies are effectively distilled into the learned utility scores themselves. Encoder-based models, being lightweight and inherently suited for regression, are thus better positioned to leverage this distilled signal, explaining their strong performance in state-of-the-art reranking systems. Further re-

search on leveraging this utility regression capabilities demonstrated by encoders while simultaneously benefiting from the comprehensive contextual understanding offered by large-context decoders would be beneficial in more complex open-domain query-based generation tasks.

Limitations

Current multi-hop datasets are predominantly fact-based. Consequently, a model trained on such data may not generalize effectively to tasks requiring high-level inferential reasoning. For example: answering questions like, "Why was the publication of The Catcher in the Rye considered provocative when it was released?". In these scenarios, the *utility* of a passage might be tied to more abstract semantic properties.

Ethics Statement

There are no ethical concerns to the best of our knowledge.

References

Yang Ba, Michelle V. Mancenido, and Rong Pan. 2024. Fill in the gaps: Model calibration and generalization with synthetic data.

George Arthur Baker, Ankush Raut, Sagi Shaier, Lawrence E Hunter, and Katharina von der Wense. 2024. Lost in the middle, and in-between: Enhancing language models' ability to reason over long contexts in multi-hop qa.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Kalervo Järvelin and Jaana Kekäläinen. 2017. Ir evaluation methods for retrieving highly relevant documents. *SIGIR Forum*, 51(2):243–250.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xiaoyan Bai. 2023. Promptrank: Unsupervised keyphrase extraction using prompt.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *Proceedings of the 7th International Conference on Learning Representations*.

Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine

Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. Openai o1 system card.

Laura Perez-Beltrachini and Mirella Lapata. 2025. Uncertainty quantification in retrieval augmented question answering.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Alireza Salemi and Hamed Zamani. 2024. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2395–2400, New York, NY, USA. Association for Computing Machinery.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop questions via single-hop question composition.

Stefan Sylvius Wagner, Maike Behrendt, Marc Ziegele, and Stefan Harmeling. 2025. The power of llm-generated synthetic data for stance detection in online political discussions.

Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Wen tau Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oğuz. 2021. Answering complex open-domain questions with multi-hop dense retrieval.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Yuanhai Xue, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. Training a utility-based retriever through shared context attribution for retrieval-augmented language models.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Are large language models good at utility judgments?

A Appendix

A.1 GPT-Annotated Utility Scores and Verification

As part of our analysis, we also evaluate the upper bound performance achievable using GPT-4o annotated utility scores to rank passages. The results, summarized in Table 2, serve as a strong validation of our training data quality and support the effectiveness of reasoning-trace-based utility scoring.

As an additional check, we manually verified approximately 150–200 examples across the HotpotQA, MuSiQue, and 2WikiMultihopQA datasets. In over 95% of cases, the LLM-generated utility scores aligned with human judgments of passage usefulness, confirming their reliability as supervision signals.

A.2 Evaluating Decoder-only Models for Utility Scoring

We conduct an auxiliary experiment by fine-tuning two decoder-only models, LLaMA 3.2 1B and LLaMA 3.1 8B (Meta, 2024), to compare their performance on this regression task against our primary encoder-only model. These models were fine-tuned using the same LLM-generated utility scores under two distinct settings: (1) Pointwise Scoring: The model predicts a utility score for each passage individually, using a `(question, passage, utility_score_k)` triplet as input. (2) Listwise Scoring: The model is presented with the question and the entire set of candidate passages simultaneously, aiming to predict individual scores with full context. Figures 2 and 4 show that providing the decoder models with full context (listwise) offers only marginal benefits over the pointwise scoring for HotpotQA and 2WikiMultiHopQA. However, for the MuSiQue dataset (Fig 3), which has a larger candidate set (20 passages), the pointwise

Dataset	P@2	R@2	F1@2	R@5	NDCG@1	NDCG@5	EM
HotpotQA	93.29	93.29	93.29	98.39	96.99	95.85	89.77
MuSiQue	87.90	79.38	83.43	84.32	93.40	86.88	82.80
2WikiMultihopQA	98.20	86.32	91.87	98.85	99.80	98.92	78.96

Table 2: Upper bound performance using GPT-annotated utility scores for ranking.

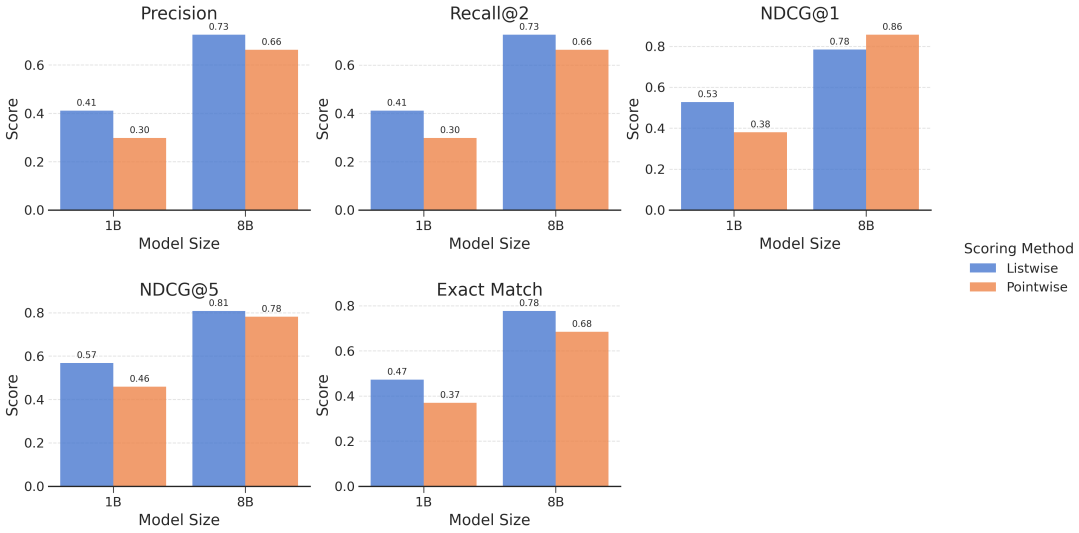

Figure 2: Performance comparison of decoder-only models (LLaMA 3.2 1B and LLaMA 3.1 8B) on the HotpotQA dataset, fine-tuned using two different methods: Pointwise and Listwise scoring

approach is significantly more effective. We hypothesize this performance degradation in the listwise setting is due to the well-documented Lost-in-the-middle (Baker et al., 2024; Liu et al., 2023) problem, where decoder models struggle to attend to all items in a long input sequence. It is also noteworthy that the absolute performance of these fine-tuned LLaMA models is mostly lower than our lightweight RoBERTa-based model (as reported in Table 1). Despite having a more constrained input during inference (scoring each passage individually), our model is better at identifying the useful passages and correctly ordering them, further underscoring the quality of our training data.

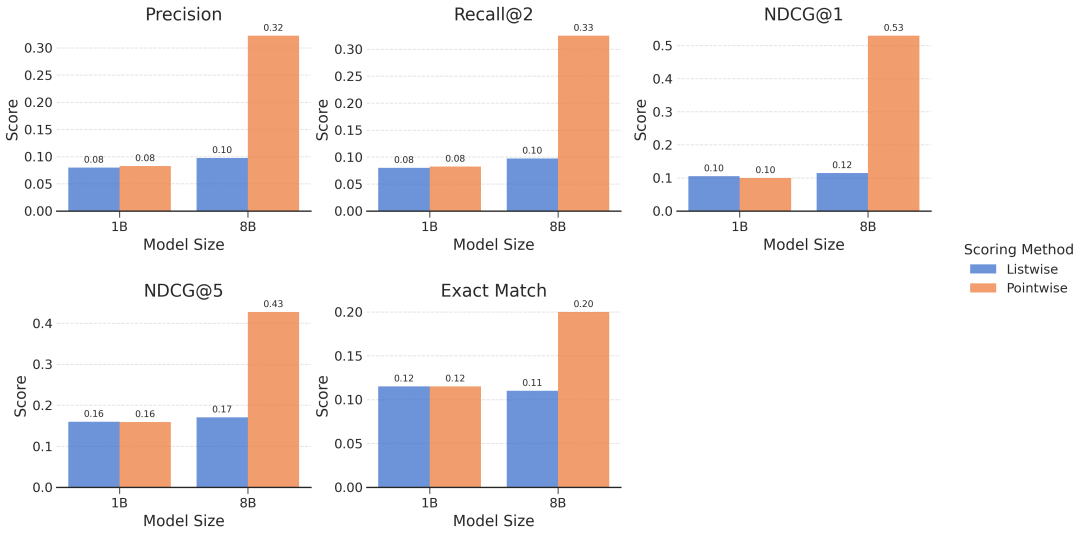

Figure 3: Performance comparison of decoder-only models (LLaMA 3.2 1B and LLaMA 3.1 8B) on the MuSiQue dataset, fine-tuned using two different methods: Pointwise and Listwise scoring

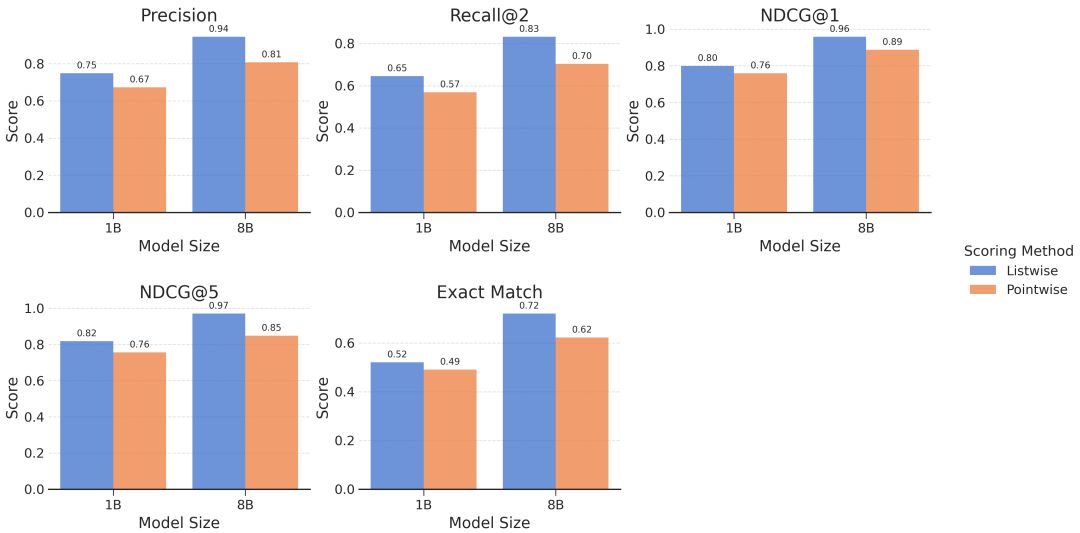

Figure 4: Performance comparison of decoder-only models (LLaMA 3.2 1B and LLaMA 3.1 8B) on the 2WikiMultiHopQA dataset, fine-tuned using two different methods: Pointwise and Listwise scoring