LimRank: Less is More for Reasoning-Intensive Information Reranking

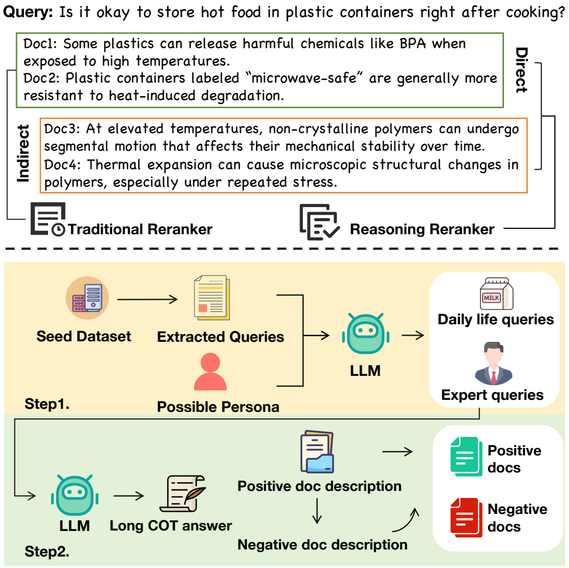
Tingyu Song* **Yilun Zhao*** **Siyue Zhang** **Chen Zhao** **Arman Cohan**

Yale NLP Lab

Abstract

Existing approaches typically rely on large-scale fine-tuning to adapt LLMs for information reranking tasks, which is computationally expensive. In this work, we demonstrate that modern LLMs can be effectively adapted using only minimal, high-quality supervision. To enable this, we design LimRank-Synthesizer, a reusable and open-source pipeline for generating diverse, challenging, and realistic reranking examples. Using this synthetic data, we fine-tune our reranker model, LimRank. We evaluate LimRank on two challenging benchmarks, *i.e.,* Bright for reasoning-intensive retrieval and FollowIR for instruction-following retrieval. Our experiments demonstrate that LimRank achieves competitive performance, while being trained on less than 5% of the data typically used in prior work. Further ablation studies demonstrate the effectiveness of LimRank-Synthesizer and the strong generalization capabilities of LimRank across downstream tasks, including scientific literature search and retrieval-augmented generation for knowledge-intensive problem solving.

yale-nlp/LimRank



Query: Is it okay to store hot food in plastic containers right after cooking?

Doc1: Some plastics can release harmful chemicals like BPA when exposed to high temperatures.
Doc2: Plastic containers labeled "microwave-safe" are generally more resistant to heat-induced degradation.

Doc3: At elevated temperatures, non-crystalline polymers can undergo segmental motion that affects their mechanical stability over time.
Doc4: Thermal expansion can cause microscopic structural changes in polymers, especially under repeated stress.

Figure 1: (Top) An illustration of reasoning-intensive reranking scenarios that demand more than surface-level semantic matching. These tasks require multi-step inference, contextual reasoning, and recognition of implicit relationships between queries and documents. (Bottom) An overview of LimRank-Synthesizer, which generates high-quality training data for reranking tasks.

1 Introduction

Recent studies (Peng et al., 2025; Zhuang et al., 2025; Chen et al., 2024) have increasingly leveraged LLMs for reranking tasks in information retrieval. While LLMs have shown effectiveness in general-purpose reranking scenarios, emerging research reveals notable limitations when these models are applied to reasoning-intensive retrieval settings (Zhang et al., 2025; Weller et al., 2025b; Shao et al., 2025; Song et al., 2025b). These limitations are not confined to performance metrics alone. Rather, they stem from the difficulty LLMs face when relevance depends on more than surface-level keyword overlap or shallow semantic similarity. In such scenarios, effective retrieval requires multi-step inference, contextual reasoning, and the ability to recognize implicit relationships between queries and documents (Su et al., 2024).

Inspired by the growing success of reasoning LLMs (Jaech et al., 2024; Guo et al., 2025), several recent studies (Weller et al., 2025b; Zhuang et al., 2025; Yan et al., 2025) have begun exploring the training of LLMs that can leverage test-time computation to improve performance in reasoning-intensive retrieval tasks. However, contemporary approaches typically rely on large-scale supervised fine-tuning. For instance, Rank1 (Weller et al., 2025b) is fine-tuned on over nearly 700K examples of DeepSeek-R1's reasoning traces, which is

*Equal Contributions. Correspondence: Yilun Zhao (yilun.zhao@yale.edu).

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 20625–20639
November 4-9, 2025 ©2025 Association for Computational Linguistics

expensive in terms of compute and data.

We hypothesize that frontier LLMs already possess considerable latent reasoning capabilities for reranking, and that these capabilities can be activated and steered using a small number of carefully curated, high-quality examples that encourage extended deliberation. This *"less is more"* approach has shown promise in other domains: for example, LIMA (Zhou et al., 2023) and LIMO (Ye et al., 2025) achieve strong performance in instruction following and complex reasoning with minimal but strategically selected examples, demonstrating that carefully curated demonstrations can effectively steer pretrained models without the need for massive fine-tuning. To our knowledge, we are the first to investigate this paradigm in reranking tasks.

We introduce LIMRANK-SYNTHESIZER, a modular and open-source pipeline for generating high-quality reranking training data through several novel design choices. LIMRANK-SYNTHESIZER is guided by three core principles: domain diversity, alignment with real-world use cases, and difficulty diversity. It generates retrieval queries paired with corresponding positive and negative passages, and employs frontier reasoning models (*i.e.,* DeepSeek-R1) to produce multi-step reasoning traces. We then apply LLM-based filtering to discard low-quality traces. Using LIMRANK-SYNTHESIZER, we generate a compact yet effective dataset of 20K examples—only 2.85% of the data used in Rank1. We fine-tune Qwen2.5-7B on this dataset to produce our reranker model, LIMRANK.

We evaluate LIMRANK on two challenging reranking tasks: (1) reasoning-intensive retrieval using BRIGHT (Su et al., 2024), and (2) instruction-following retrieval using FOLLOWIR (Weller et al., 2025a). LIMRANK achieves the nDCG@10 score of 28.0% on BRIGHT and p-MRR score of 1.2 on FOLLOWIR, representing the best performance among models with 7B-level parameters. To better understand the strengths and limitations of LIMRANK, we conduct an in-depth human evaluation, revealing that LIMRANK excels particularly in settings that require multi-hop reasoning, subtle instruction disambiguation, and context-sensitive reranking. Additionally, we conduct extensive ablation studies on LIMRANK-SYNTHESIZER and show that each component of our guidelines is essential. We also train different model variants using synthetic IR datasets of the same size (*i.e.,* 20K). Models trained with our data consistently outperform those trained with other synthetic IR datasets used by RANK1, Promptriever, and ReasonIR.

To assess LIMRANK's practical utility, we further deploy and evaluate it in two real-world-inspired tasks: (1) Scientific Literature Search on the LitSearch dataset (Ajith et al., 2024), and (2) Retrieval-Augmented Generation (RAG) on the GPQA benchmark (Rein et al., 2024). Compared to the previous state-of-the-art, Rank1-7B, LIMRANK demonstrates strong generalization, achieving 30.3% accuracy on GPQA (vs. 28.3%) and a competitive 60.1% Recall@5 on LitSearch (vs. 60.8%). , demonstrating its effectiveness as a plug-in reranker in real-world systems.

2 Related Works

Reasoning-intensive Information Retrieval (Su et al., 2024) has gained increasing attention due to its practical relevance in real-world scenarios. Some works, such as RANK1 (Weller et al., 2025b) and ReasonIR (Shao et al., 2025), address this challenge by fine-tuning models on large-scale, reasoning-focused retrieval datasets to enhance the performance of rerankers and retrievers, respectively. Recently, other works (Ji et al., 2025; Abdallah et al., 2025; Liu et al., 2025) have proposed complex training strategies using synthetic data. However, the large data volume required by these methods leads to high training costs, and the training strategies themselves introduces extra training cost. Some other works (Jin et al., 2025; Song et al., 2025a; Zhuang et al., 2025) employs reinforcement learning to equip retrievers and rerankers with reasoning capabilities. While effective, these methods often entail significant computational, data, and engineering overhead. Recent studies (Zhou et al., 2023; Ye et al., 2025) suggest that modern LLMs possess substantial latent reasoning abilities across a range of tasks, which can be efficiently activated and enhanced using small amounts of task-specific training data. To support this, various data selection methods (Xia et al., 2024; Liu et al., 2024) have been proposed and shown to be effective. To our best knowledge, our work is the first to apply such *"less is more"* exploration to reranking tasks.

3 LIMRANK-SYNTHESIZER

3.1 Data Curation Guidelines

LIMRANK-SYNTHESIZER is designed to unlock the LLM's latent potential by minimizing the number of training examples while maximizing their informativeness and depth of reasoning. We estab-

Element	Data Generation Guideline
Query	***Domain Diversity:*** Queries should not limited to everyday contexts but also span specialized domains such as finance, law, and healthcare.
	Alignment with Daily: Life Queries vary in complexity to mirror real-world use cases. Simple queries reflect straightforward needs, while complex ones mimic intricate scenarios.
	Difficulty Diversity: The dataset should contain simple queries with challenging ones. Hard queries incorporate patterns like instruction-following, multi-step reasoning, and so on.
Reasoning Chain	***Well-Structured Reasoning:*** Reasoning chains should be organized logically to enhance inference efficiency. Clear step-by-step processes help the model navigate solutions systematically during search operations.
	Human Verification: Reasoning chain undergoes human review to validate accuracy, coherence, and alignment with the query's intent. This ensures robust, error-resistant logic.
	Adaptive Analysis: For simple queries, reasoning focuses on explicit connections between the query and passage. For complex queries, chains prioritize identifying implicit relationships, resolving ambiguities, and addressing layered contextual demands
Passage	***Hard Negatives:*** Hard negative passages with subtle alterations can help train the model in discerning fine-grained distinctions.
	Complexity Diversity: Passages vary in structure and content difficulty, including lengthy texts, analytical data (e.g., statistical reports), or domain-specific jargon.

Table 1: Data generation guideline. We encourage the query to be diverse. Consequently, the reasoning chain will be more diverse. We believe this can activate LLM potential more easily.

lish a set of guidelines to govern the construction of training data, as detailed in Table 1.

3.2 Reasoning-intensive Data Generation

Commonly used training datasets for IR tasks (*i.e.,* MS MARCO) are limited in complexity. The queries in these datasets are typically straightforward, which can constrain model performance on reasoning-intensive retrieval tasks. Recently various data synthesis methods (Chen et al., 2023; Shen et al., 2025) are proposed in retrieval and fact checking domains. Some works (Almeida and Matos, 2024; Sinha, 2025) focus on constructing the dataset with higher quality or with better efficiency. Recent studies such as ReasonRank (Liu et al., 2025) and DIVER (Long et al., 2025) have demonstrated strong results on reasoning-intensive tasks using specialized synthetic data. A limitation of these methods is their dependency on large-scale data aggregation from varied sources. This work, therefore, focuses on creating a straightforward and easy-to-implement pipeline. We propose LimRank-Synthesizer, a bottom-up data synthesis pipeline designed to generate high-quality training data. As illustrated in Figure 1, LimRank-Synthesizer uses MS MARCO as the seed dataset and generates enhanced data based on its simple queries, guided by our designed guideline discussed earlier. We detail the query generation and passage generation processes as follows:

Query Generation. Recognizing that individuals from different backgrounds focus on different aspects of a topic, we use a variety of personas to ensure query diversity. We first randomly sample several personas from PersonaHub (Ge et al., 2025) and use a query sampled from the seed dataset to prompt the LLM to generate a persona that suits the context. Given the same query and a specific persona, we then prompt an LLM (*i.e.,* GPT-4o) to generate two types of augmented queries: one situated in a daily-life context and the other in an expert-domain context. For daily-life scenarios, we encourage the LLM to rewrite queries such that they involve more complex or nuanced real-world situations, that may challenge everyday human reasoning. For expert-domain scenarios, we focus on questions requiring domain-specific knowledge. These include queries seeking evidence, challenging a claim, or reasoning about complex professional issues. All generated queries are derived from the original MS MARCO queries, using carefully designed prompts detailed in Appendix A.1.

Positive and Negative Passage Generation. We construct passages that are both directly and indirectly related to the query. To achieve this, we employ the CoT prompting technique to guide the LLM in describing the intermediate materials needed to solve the problem. The final positive passage is then generated based on these descriptions. As prior work (Wang et al., 2023) has shown, hard

	StackExchange							Coding		Theroem-based			Avg
	Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	Aops	TheoQ.	TheoT.	
BM25s	18.2	27.9	16.4	13.4	10.9	16.3	16.1	**24.7**	4.3	6.5	7.3	2.1	13.7
Monot5-3B	37.9	45.7	24.1	34.3	17.6	24.1	25.1	18.2	21.5	4.9	19.8	21.3	24.5
RankZephyr-7B	21.9	23.7	14.4	10.3	7.6	13.7	16.6	6.5	24.7	**6.8**	2.0	7.3	13.0
RankGPT	33.8	34.2	16.7	27.0	**23.3**	**27.7**	11.1	15.6	3.4	1.2	8.6	0.2	17.0
RankLLaMA-7B	40.9	44.0	23.6	35.6	21.1	23.5	25.6	16.5	23.3	3.7	14.2	14.7	23.9
Rank-R1	26.8	24.8	17.9	22.1	17.4	10.3	21.1	4.4	15.6	3.3	10.4	5.9	15.0
Qwen2.5-7B	50.7	38.5	21.0	30.0	16.0	21.1	22.9	18.0	15.3	3.7	17.4	16.3	22.6
Rank1-7B	48.8	36.7	20.8	35.0	22.0	18.7	**36.2**	12.7	**31.2**	6.3	**23.7**	**37.8**	27.5
LimRank	**52.5**	**49.8**	**25.4**	**43.1**	22.4	25.7	27.9	17.5	20.2	4.7	20.0	27.3	**28.0**

Table 2: nDCG@10 results on Bright. All methods rerank the top-100 documents retrieved by BM25. Results for BM25, RankZephyr-7B, RankGPT4, and RankR1 are copied from Rank-R1.

negative passages are crucial for effective training. To create such negatives, we provide the LLM with the query and the associated positive descriptions, and then prompt it to generate corresponding hard negative descriptions. And these negative descriptions are used in the same way as positive's to generate negative passages. The prompts used for both positive and negative generation are included in Appendix A.2.

3.3 Data Filtering

Following Rank1, we ask DeepSeek-R1 to judge the relevance between query and the passage. If the judgment is the same as we generated previously, then we collect this dataset into our final dataset and filter the rest of them. Finally, we collected 10,282 (query, passage) pairs.

4 Experiments

This section presents our main results and ablation studies. We provide details on the experimental setup, including evaluation setting, baseline systems, and implementation specifics, in Appendix B.

4.1 Main Results

Bright. We present results on Bright in Table 2. Compared to Rank1, which adopts a similar approach to LimRank for training a pointwise reranker, LimRank surpasses its performance and achieves state-of-the-art results on this task. Although the improvement is modest, it is noteworthy that our model is trained with significantly less data than Rank1 (20K vs. 6 million examples), This not only reduces training cost but also partially alleviates the efficiency concerns associated with Rank1. Moreover, when compared to other models, LimRank shows a clear performance advantage. Listwise models like RankGPT4 and RankerZephyr are not explicitly trained for reasoning-intensive scenarios, which may account for their lower performance. The setwise reranker Rank-R1, trained via RL, also falls behind LimRank, despite setwise rerankers generally having structural advantages over pointwise models. This suggests that either the reward function or the training data used in Rank-R1 may require more careful design and curation.

FollowIR. Instruction-following has garnered increasing attention in the IR community. As shown in Table 3, LimRank achieves state-of-the-art performance on FollowIR. Notably, LimRank outperforms Rank1 while using substantially less training data, which may be attributed to the increased diversity and complexity of the synthesized training data. This supports the conclusion that a smaller amount of high-quality data can more effectively activate the model's instruction-following capabilities. However, the performance gap between LimRank and FollowIR-7B remains relatively small. This implies that test-time scaling alone may not be sufficient for enhancing instruction-following performance, highlighting the importance of carefully designed training data even for large models.

4.2 Error Analysis

LimRank fall short in some cases, we conclude them as follows: (1) **Queries requiring directly relevant passages.** When the query clearly implies a need for directly relevant content, LimRank typically gives similar scores to both direct and indirect relevant passages. (2) **Queries with very few relevant passages.** In cases where only one or two golden passages exist, LimRank exhibits uncertainty during the relevance judgment process and

Model	FOLLOWIR Score	 p-MRR	LitSearch Recall@5	GPQA Acc
Qwen2.5-7B	11.6	-0.8	49.5	26.3
RankLLaMA-7B	15.5	0.0	59.0	28.3
FollowIR-7B	15.6	0.3	57.5	**30.3**
RANK1-7B	16.8	0.3	**60.8**	28.3
LIMRANK	**17.5**	**1.2**	60.1	**30.3**

Table 3: Average performance on relevant tasks. Detailed results can be found in Appendix B.

is more likely to fail. These observations suggest a potential direction for future work: developing an adaptive reranker capable of distinguishing query intent and assigning relevance scores with greater contextual awareness. And we provide a detailed case study in Appendix C.2.

4.3 Experiments on Real-world Tasks

To assess the practicality of LIMRANK in real-world applications, we evaluate it on two downstream tasks. The results are shown in Table 3. LitSearch (Ajith et al., 2024) is a retrieval dataset containing natural queries in the scientific literature domain. For the *specific* question type in LitSearch—where no more than five papers are relevant—LIMRANK achieves a Recall@5 of 60.1%, which is comparable to the 60.8% achieved by RANK1, indicating only a negligible performance gap. GPQA (Rein et al., 2024) is a challenging benchmark covering the domains of biology, physics, and chemistry. To assess real-world usefulness, we evaluate LIMRANK on this dataset within a RAG setup. LIMRANK achieves state-of-the-art performance with an accuracy of 30.3%, outperforming RANK1 's 28.3%. Overall, LIMRANK demonstrates comparable performance across downstream tasks within less data involved in training. Detailed experimental settings and analysis are provided in Appendix B.5.

4.4 Analysis of Synthetic Training Data

To validate our methodology, we assess both the individual components of our guidelines and the overall quality of our synthetic pipeline.

Dataset Design Components. From Table 4, we can find that each component of our guidelines is essential. Simple queries from RANK1 are insufficient, while daily life queries greatly boost reasoning and RAG performance. Adding long reasoning traces improves both instruction following and RAG, highlighting the importance of query-adaptive reasoning.

	BRIGHT	FollowIR	LitSearch	GPQA
LIMRANK (Full Set)	**28.0**	**1.19**	**60.1**	**30.3**
Daily-life Queries	27.6	0.39	59.6	26.8
Expert Domain Queries	26.7	0.85	59.1	26.8
Short Reasoning Trace	27.0	0.72	59.9	27.3
Long Reasoning Trace	26.7	1.17	59.3	27.3

Table 4: Performance comparison across different configurations of our multi-level guideline. LIMRANK refers to the full version.

	BRIGHT	FollowIR	LitSearch	GPQA
Promptriever	25.5	0.96	59.8	28.8
ReasonIR	25.6	0.21	**60.4**	26.8
LIMRANK	**28.0**	**1.19**	60.1	**30.3**

Table 5: Performance comparison between different synthesis datasets.

Training Dataset Size. As shown in Table 3, LIMRANK achieves strong performance on both reranking and real-world application tasks, despite being trained on far less data. This raises the question of whether the scaling law still applies in this context. Although some prior works (Ye et al., 2025) demonstrate that even small datasets can yield strong results, we observe that performance plateaus when the data size becomes too limited.

Effectiveness of LIMRANK-SYNTHESIZER. To further examine the effectiveness of our synthesized data, we trained several reranker variants using different synthetic IR datasets of the same size (20K examples), including RANK1, Promptriever, ReasonIR, and LIMRANK. As shown in Table 5, the variant trained with LimRank consistently outperforms those trained on the other synthetic datasets, demonstrating its effectiveness for training rerankers in low-resource settings.

5 Conclusion

We propose LIMRANK-SYNTHESIZER, a data synthesis pipeline that generates high-quality IR training data. Trained with far fewer but higher-quality examples, LIMRANK achieves state-of-the-art results on the reasoning and instruction-following benchmarks, and demonstrates strong performance on downstream tasks. Overall, LIMRANK provides empirical support for the *"Less is More Hypothesis"* in the IR field, offering a cost-effective alternative to data-intensive approaches.

Limitations

In this section, we highlight two limitations of our work, each of which offers promising directions for future research. First, our data selection method is both naive and costly. As described earlier, we rely on human verification during the data generation process and apply basic filtering techniques. While effective, this approach is resource-intensive and may not scale efficiently. Developing automated verification strategies or more sophisticated filtering methods could significantly improve efficiency and reproducibility in future work. Second, we apply our synthesized training data only to pointwise rerankers. Although our method shows strong performance in this setting, it remains unexplored how well the data would transfer to other architectures, such as listwise or setwise rerankers. Extending our pipeline to support these paradigms may further enhance generalization and flexibility. These limitations suggest several directions for future work.

References

Abdelrahman Abdallah, Jamshid Mozafari, Bhawna Piryani, and Adam Jatowt. 2025. Dear: Dual-stage document reranking with reasoning agents via llm distillation. *arXiv preprint arXiv:2508.16998*.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. 2024. Litsearch: A retrieval benchmark for scientific literature search.

Tiago Almeida and Sérgio Matos. 2024. Exploring efficient zero-shot synthetic dataset generation for information retrieval. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1214–1231.

Mingda Chen, Xilun Chen, and Wen-tau Yih. 2023. Few-shot data synthesis for open domain multi-hop question answering. *arXiv preprint arXiv:2305.13691*.

Yiqun Chen, Qi Liu, Yi Zhang, Weiwei Sun, Daiting Shi, Jiaxin Mao, and Dawei Yin. 2024. Tourrank: Utilizing large language models for documents ranking with a tournament-inspired strategy. *arXiv preprint arXiv:2406.11678*.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling synthetic data creation with 1,000,000,000 personas.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Yuelyu Ji, Zhuochun Li, Rui Meng, and Daqing He. 2025. Reason-to-rank: Distilling direct and comparative reasoning from large language models for document reranking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2320–2329.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. 2025. Reasonrank: Empowering passage ranking with strong reasoning ability. *arXiv preprint arXiv:2508.07050*.

Zikang Liu, Kun Zhou, Wayne Xin Zhao, Dawei Gao, Yaliang Li, and Ji-Rong Wen. 2024. Less is More: High-value Data Selection for Visual Instruction Tuning.

Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yue Shen, Jian Wang, Peng Wei, Jinjie Gu, and Jiahai Wang. 2025. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Zhiyuan Peng, Ting ruen Wei, Tingyu Song, Yilun Zhao, and Yi Fang. 2025. Efficiency-effectiveness reranking flops for llm-based rerankers.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen tau Yih, Pang Wei Koh, and Luke Zettlemoyer. 2025. Reasonir: Training retrievers for reasoning tasks.

Zhiyu Shen, Jiyuan Liu, Yunhe Pang, and Yanghui Rao. 2025. Hopweaver: Synthesizing authentic multi-hop questions across text corpora. *arXiv preprint arXiv:2505.15087*.

Aarush Sinha. 2025. Don't retrieve, generate: Prompting llms for synthetic training data in dense retrieval. *arXiv preprint arXiv:2504.21015*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025a. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Tingyu Song, Guo Gan, Mingsheng Shang, and Yilun Zhao. 2025b. IFIR: A comprehensive benchmark for evaluating instruction-following in expert-domain information retrieval. In *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)*, pages 10186–10204, Albuquerque, New Mexico. Association for Computational Linguistics.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2024. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2025a. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In *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)*, pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. 2025b. Rank1: Test-Time Compute for Reranking in Information Retrieval.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting Influential Data for Targeted Instruction Tuning.

Ruiran Yan, Zheng Liu, and Defu Lian. 2025. O1 embedder: Let retrievers think before action.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. LIMO: Less is More for Reasoning.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. 2025. Diffusion vs. autoregressive language models: A text embedding perspective.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. 2025. Rank-R1: Enhancing Reasoning in LLM-based Document Rerankers via Reinforcement Learning.

A Prompts

A.1 Prompt for Query Expansion

We provide the prompts used for query generation as follows. First, we incorporate personas from PersonaHub (Ge et al., 2025) to personalize the queries. Then, as described previously, we expand each query into two types: one set reflecting real-life scenarios and the other targeting expert domains. The prompt used for persona integration is shown in Figure 2. And we random sample personas from PersonaHub each time. We use Figure 3 for real-life scenario expansion and Figure 4 for expert domain expansion.

A.2 Prompt for Passage Generation

As mentioned earlier, we first use Figure 5 to prompt the LLM (*i.e.,* GPT-4) to solve the query using a detailed chain-of-thought (CoT) reasoning process. We then extract relevant passage descriptions using Figure 6. To ensure that the negative

```
Based on the query, please generate a persona similar to the example personas below.

Query: [FILL_QUERY_HERE]

Here are some example personas for reference: [FILL_EXAMPLES_HERE]

Please provide only the personalized information relevant to the query, without any additional content.
```

Figure 2: Prompt for query expansion in daily life scenario.

```
Given a query and a persona, please concrete the query by incorporating the persona. Query: [FILL_QUERY_HERE]
Persona: [FILL_PERSONA_HERE]

Here are some requirements for the concretized query:
1. The new query should not be a direct combination of the query and the persona.
2. It should be a newly formed query influenced by the persona.
3. Don't asking too many questions. Please focues only 1-2 questions.

And you should provide a context or scenario related to the new query.
Here are some requirements for the scenario:
1. The context and scenario should be narrated in first person or narrate some facts.
2. The context or scenario can be the background or the situation which lead to the new query.
3. The context or scenario should be within a main line and in a logical order.
4. The context or scenario should have some controversial points leading to the new query. But you should not include "controversial" and similar words in the context or scenario.

You should return with in the following format:
“‘json{
"query" : "the new query",
"scenario" : "the scenario or context related to the new query"
}“‘
```

Figure 3: Prompt for query expansion in daily life scenario.

passages are sufficiently challenging, we provide the LLM with the query and the positive passage description, and use Figure 7 to generate hard negative descriptions. Finally, we generate the corresponding positive and negative passages using Figure 8 for each given descriptions.

A.3 Prompt for R1

The prompt we used for DeepSeek-R1 to get the reasoning chain and the final relevance judgement is the same as Rank1. It is shown in Figure 9.

B Experiment Settings and Details

B.1 Main Experiments Settings

LimRank Settings. We adopt Qwen2.5-7B from the Qwen2.5 (Yang et al., 2025) series as our backbone model. The training data consists of two components: 14,000 examples from MS MARCO used in Rank1 (Weller et al., 2025b), and 6,000 examples generated by LimRank-Synthesizer. We ensure that positive and negative passages are balanced within the training dataset. Additional training details are provided in Appendix B.2.

Datasets. We evaluate LimRank on two key abilities required for reranking: reasoning and instruction following. For reasoning, we benchmark performance across all subtasks of Bright. we retrieve the top-100 documents using BM25 and compute nDCG@10 on the reranked results. To assess instruction-following ability, we evaluate on FollowIR. Unlike the original FollowIR setup, we retrieve only the top-100 documents using BM25 instead of the top-1000. Except the number of documents to rerank, we follow the same evaluation settings as in FollowIR. We mainly calculate top-5 documents' scores in nDCG, MAP and p-MRR. We provide detailed results in Appendix B.4. We further provide an analysis on the downstream task LitSearch (Ajith et al., 2024) and GPQA (Rein et al., 2024) to assess real-world applicability. For LitSearch, we mainly benchmark LimRank on the “Specific” type questions in LitSearch and calculate Recall@5. For the GPQA RAG experiment, we use the data store constructed in ReasonIR (Shao et al., 2025) as the corpus. And we choose the diomand subset for test. We ask the reranker to rerank the top-100 documents retrieved from BM25. We use

Given a query, please make a new query.

Query: [FILL_QUERY_HERE]

Here are some requirements for the new query:
1. The new query should be in the same field as the original query. But the new query should be a different topic. For example the original query is in the history field, the new query should be in the history field but a different topic.
2. The new query should be a more professional query that can include some professional terms or jargons.
3. The new query should be a more complex query. But it should mainly focus on 1-2 points. But 1 point is better.
4. The new query should be expanded like one's thinking process. It should show one's exploration or thinking in an expert-domain and final lead to the new query.
5. The new query can include statistical number or formula, and some research papers or news.

So what you should do is:
1. In-breathe the query to another topic but in the same field. And the query should be a more professional query. The in-breathing query should focus mainly on 1-2 small points.
2. Expand the new query like one's thinking or exploration process. The person can be influenced by some papers, politics, or some events. And someone maybe confused by the conflict or contradiction between their recognition and the facts. These should be included in the newly generated query.

For example,
1. As far as I understand, overfitting in machine learning happens when a model learns the training data too well, including its noise, and thus performs poorly on unseen data. Regularization techniques like dropout or L2 loss are supposed to prevent overfitting by penalizing complexity.
But in practice, large models trained on huge datasets—like GPT-style models—seem to avoid overfitting even when they're massively overparameterized. Why don't they overfit despite having more parameters than training samples?
If it's due to the scale of data, then shouldn't smaller models trained on large data also generalize as well? And if it's about optimization dynamics, why do simple regularizers help at all in small-scale settings?
2. The fall of the Western Roman Empire in 476 AD is commonly seen as the end of ancient Rome and the beginning of the Middle Ages. The deposition of Romulus Augustulus by Odoacer is often cited as the decisive event.
But Roman institutions, language, and even laws continued to exist in various forms across Europe for centuries—especially in the Byzantine Empire, which still called itself Roman. If Rome "fell," why did so much of its structure persist?
If the fall of Rome was a definitive collapse, why didn't it lead to a more immediate or total cultural break? And if it was more of a transformation, should we even think of it as a "fall" at all?

Please only provide the new query, without any additional content.

Figure 4: Prompt for query expansion in daily life scenario.

Given a query, please analyze and answer this question, providing the reasoning process for each step. Respond point by point. The response should include the materials needed for the analysis. Please think step by step.

Query: [FILL_QUERY_HERE]

Please return as detailed a response as possible, including the reasoning process for each step.
...

Figure 5: Prompt for problem solving.

Given a passage, please extract the materials described in the passage.
Here is the passage: [FILL_PASSAGE_HERE]

You need to extract the material description. The number of the material description should be in range 3-7.
What you extracted need to cover different aspects of the origin passage. And please return in the following format:
1. [extracted material description 1]
2. [extracted material description 2]
...

Figure 6: Prompt for materical description extraction.

You are assigned a task in the text information retrieval field. Given a query, positive passage descriptions, you need to generate negative passage descriptions.

Query: [FILL_QUERY_HERE]
Positive passage descriptions: [POSITIVE_PASSAGE_DESCRIPTIONS_HERE]

The negative passage descriptions should follow the following guidelines:
1. The negative passage description can be partially related to the query's topic.
2. The negative passage description can be used to generate hard negative passages.
3. You should consider the given positive passage descriptions. The negative passage description should be different from the positive passage descriptions.

You can return 1-5 negative passage descriptions. And you should return in a list format as follows:
1. [negative passage description 1]
2. [negative passage description 2]
...

Figure 7: Prompt for query expansion in daily life scenario.

Given a material description, please generate a passage that satisfies the material description.

Here is the material description: [FILL_MATERIAL_DESCRIPTION]

The generated passage should satisfy the following requirements:
- Satisfying the material description.
- Include the number or analysis if the material description requiring analysis or calculation.
- Include the policy or principle if the material description requiring policy or principle.
- Include the legal cases or laws if the material description requiring legal cases or laws.
- Include the examples or evidence if the material description requiring examples or evidence.

Anyway, please generate a passage that satisfies the material description.
Please only provide the passage, without any additional content.

Figure 8: Prompt for passage generation.

Determine if the following passage is relevant to the query. Answer only with 'true' or 'false'.
Query: [FILL_QUERY_HERE]
Passage: [FILL_PASSAGE_HERE]

Figure 9: Prompt for Deepseek-R1 judgement.

the Qwen2.5-7B as the reader and set topk to be three. We provide a further analysis of the RAG experiment in Appendix B.5.

Baselines. We use BM25 (Robertson et al., 2009) as the initial retriever to obtain the top-100 candidate documents. For a fair comparison, we primarily evaluate LIMRANK against models with similar parameter sizes. We include listwise rerankers such as RankerZephyr-7B (Pradeep et al., 2023) and RankGPT4 (Sun et al., 2023), as well as the setwise reranker Rank-R1 (Zhuang et al., 2025). Additionally, we include MonoT5 (Nogueira et al., 2020), a widely adopted pointwise reranking model, as a classic baseline.

B.2 Training and Hyperparameter Details

We primarily follow the experimental settings from RANK1. All experiments are conducted on machines with 2×80GB H100 GPUs. Models are trained for 5 epochs, with each training run taking less than 2 hours. We use LLaMA-Factory for LoRA fine-tuning on all parameters, with a rank of 32 and alpha set to 64. The learning rate is set to 6e-5, and the batch size is 128.

B.3 Ablation Experiments

Training Dataset Size We describe the settings of our data scaling experiments, with results shown in Table 6. For the 2K setup, we use 1,000 samples from the RANK1 training data and 1,000 samples from LIMRANK-SYNTHESIZER. For the 10K setup, we use 6,000 samples from RANK1 and 4,000 from LIMRANK-SYNTHESIZER. For the 20K setup, we use 14,000 samples from RANK1 and 6,000 from LIMRANK-SYNTHESIZER. All models in these experiments are trained under the same settings.

Dataset Design Components. We conduct extra ablation experiments in rebuttal stage to validate that each point in the guideline is necessary. We choose variables like, simple(hard) queries, daily(expert) queries, short(long) reasoning trace. All experiments are conducted with the same 20k training data size. We find that: (1) Simple queries from RANK1 is not enough. (2) Daily life queries can bring huge improvement in reasoning-intensive tasks and RAG scenarios. But it may fails on instruction-following instructions. (3) Long reasoning trace can benefit instruction following abilities and performance in the RAG performance. (4) This emphasizes the importance of including reasoning traces with query-adaptive lengths in the training data. And we will provide a detailed analysis in the revised manuscripts.

Effectiveness of Our Dataset To further examine the effectiveness of LimRank data, we trained several reranker variants using different synthetic IR datasets of the same size, including RANK1, Promptriever, ReasonIR, and LIMRANK. All models were trained under identical settings to ensure a fair comparison. As shown in the results, the variant trained with LIMRANK consistently outperforms the other synthetic datasets, demonstrating its effectiveness compared to other methods in training rerankers in a low-resource setting.

Importance of Reasoning Traces. Under the 20K-example setting, we evaluate both the presence and absence of reasoning traces across LIMRANK, ReasonIR, and RANK1. We find that: (1) Training RANK1 with reasoning traces leads to improved performance. (2) LimRank outperforms both ReasonIR and RANK1 in settings with and without reasoning traces.

B.4 FollowIR Results

We provide the details of FOLLOWIR results in Table 7.

B.5 Downstream Task Details

LitSearch Experiment The LitSearch dataset includes two types of questions: “broaden” and “specificity”. In this work, we focus on the “specificity” questions, which require retrieving the top five most suitable documents. This setting imposes a higher standard of precision.

Retrieval-Augmented Generation (RAG) Experiment For the RAG experiments, we use diamond subset of GPQA as the question-answering benchmark. We use the filtered datastore version from ReasonIR as the retrieval corpus. We rerank the top-20 documents and select the top-3 to serve as the input to the generation model.

C Examples

C.1 LIMRANK-SYNTHESIZER Examples

We provide extra examples of LIMRANK-SYNTHESIZER. The query examples are shown in Table 8. And the passage examples are provided in Table 9. Provided passages are corresponding to the first query in Table 8.

	StackExchange							Coding		Theroem-based			Avg
	Bio.	Earth.	Econ.	Psy.	Rob.	Stack.	Sus.	Leet.	Pony	Aops	TheoQ.	TheoT.	
LIMRANK-2k	47.8	43.0	24.2	35.1	19.8	23.0	26.0	**17.8**	16.6	3.9	17.9	24.8	25.0
LIMRANK-10k	50.5	45.3	24.7	37.6	20.5	**26.3**	**28.2**	17.5	16.6	4.3	19.6	**28.0**	26.6
LIMRANK-20k	**52.5**	**49.8**	**25.4**	**43.1**	**22.4**	25.7	27.9	17.5	**20.2**	**4.7**	**20.0**	27.3	**28.0**

Table 6: Scaling data quantity influences. nDCG@10 results on BRIGHT.

Model	Robust04		News21		Core17		Avg	
	MAP	p-MRR	nDCG	p-MRR	MAP	p-MRR	Score	p-MRR
Qwen2.5-7B	7.7	0.4	24.2	-0.9	2.8	-1.9	11.6	-0.8
MonoT5-3B	10.6	0.2	30.0	0.0	4.6	-0.1	15.1	0.0
RankLLaMA-7B	9.8	0.0	32.4	0.0	4.3	0.0	15.5	0.0
FollowIR-7B	9.7	-0.0	32.8	0.3	4.1	0.5	15.6	0.3
Rank1-7B	11.4	1.6	34.0	-0.3	4.9	-0.3	16.8	0.3
LIMRANK	**11.8**	**1.7**	**35.7**	**1.0**	**4.9**	**0.8**	**17.5**	**1.2**

Table 7: Total results for FOLLOWIR. The score is measured by MAP@5, nDCG@5 and p-MRR on top-100 documents retrieved by BM25.

C.2 Error Analysis and Case Study

We compare the output from LIMRANK and RANK1. As shown in Figure 10, LIMRANK is able to identify indirect relevance between the query and the passage, whereas RANK1 marks the same passage as irrelevant. This may be attributed to the fact that our training data includes examples of indirect relevance, rather than relying solely on golden passages from traditional datasets. In addition, We provide examples in Figure 11 and Figure 12 to demonstrate LIMRANK's capability to distinguish fine-grained instruction variations. These cases collectively highlight the reasoning and instruction-following abilities exhibited by LIMRANK.

[LIMRANK Output]:
... But since the query is about making a fake signal, the passage does support the resolution by explaining the existing technology that does that. So, **even though it's not a direct answer, it's foundational and relevant.** Therefore, the answer should be true.

[RANK1 Output]:
... Even though both topics relate to the auditory system, the passage doesn't provide information on how to create fake signals. It's more about the medical device implantation rather than the technical or experimental methods to produce and deliver such signals. **Therefore, the passage doesn't address the query's core concern about generating fake auditory signals, making it irrelevant.**

Figure 10: Output comparison between LIMRANK and RANK1.

ID	Type	Original Query	Persona	Query
1	Daily Query	what did komodo dragons do	A herpetologist specializing in the behavior and ecology of Komodo dragons, dedicated to understanding their role in their natural habitat and their interactions with other species and the environment.	As a herpetologist focusing on Komodo dragons, I've spent years observing how these fascinating creatures navigate their natural environment. While conducting fieldwork on Komodo Island, I noticed that some researchers argue that Komodo dragons primarily rely on their size and strength to dominate their territory, while others suggest that their sophisticated hunting strategies and interactions with other predators contribute more significantly. This discrepancy in understanding has led me to question the specific behaviors that shape their interactions with both prey and potential competitors in their ecosystem.
2	Daily Query	what is a good glucose number	A diabetes educator specializing in glucose management, dedicated to helping individuals understand optimal blood sugar levels and their impact on overall health.	What are the optimal blood sugar targets I should aim for to effectively manage glucose levels and avoid potential health issues?
3	Expert Query	what is driverquery.exe	A computer systems analyst with a focus on Windows operating system utilities, particularly interested in exploring and educating others about built-in tools like driverquery.exe and how they can be used to optimize system performance and troubleshoot hardware issues.	As far as I am aware, Windows Management Instrumentation (WMI) is a crucial infrastructure for managing data and operations on Windows-based operating systems, providing a standardized way to access system information and resources. It is widely used for system administration, monitoring, and automation tasks. However, WMI also poses security concerns due to its capabilities, such as remote access and execution. Given the increasing sophistication of cybersecurity threats, how do organizations effectively balance the need for WMI usage in system management and automation with the potential security risks it poses? Are there specific security frameworks or best practices in place to mitigate these risks while maintaining system efficiency? Additionally, how do recent developments in cybersecurity research address the vulnerabilities associated with WMI, and what role do statistics on WMI-related security incidents play in shaping these strategies?
4	Expert Query	how much does it cost to build a bbq island	A DIY enthusiast and outdoor living specialist who is passionate about designing and constructing customized outdoor spaces, with a particular interest in building functional and aesthetically pleasing BBQ islands. They are focused on understanding the costs, materials, and design trends involved in creating the perfect outdoor cooking area.	When designing an outdoor kitchen space, especially a high-end BBQ island, the choice of materials can significantly influence both the functionality and longevity of the structure. While traditional materials like concrete and stainless steel are popular, there is an emerging interest in more sustainable and innovative materials such as recycled composite decking or Corian. However, these materials often come with a different set of maintenance requirements and environmental impacts. How do these alternative materials compare to traditional ones in terms of thermal resistance and durability, particularly when exposed to extreme weather conditions? Furthermore, considering the environmental impact, is the initial higher cost of these eco-friendly materials offset by their durability and reduced maintenance over time? Are there any recent studies or data supporting these claims, possibly influencing trends in the industry towards sustainable building practices?

Table 8: Query examples. The examples are generated by LIMRANK-SYNTHESIZER.

ID	Type	Materical Description	Passage
1	Direct Positive Document	Information from wildlife documentaries detailing Komodo dragons' hunting and feeding behaviors, including their ambush tactics and diet.	Komodo dragons, the apex predators of their island habitats, exhibit fascinating hunting and feeding behaviors that have been meticulously documented in wildlife documentaries. These formidable reptiles employ ambush tactics, using their stealth and patience to great effect. Often lying in wait along animal trails or near water sources, a Komodo dragon remains motionless, blending seamlessly into the environment with its mottled skin until an unsuspecting prey comes within striking distance. Once the prey is close enough, the dragon launches a swift and powerful attack, using its muscular legs to propel itself forward and its strong, serrated teeth to deliver a fatal bite. The diet of a Komodo dragon is diverse, reflecting its opportunistic nature. They primarily feed on large mammals, such as deer and wild boar, but they are also known to consume smaller prey like birds, snakes, and even carrion. Documentaries often highlight their remarkable ability to consume up to 80% of their body weight in a single meal, showcasing their dominance as the largest living species of lizard. Despite their size and power, Komodo dragons are not entirely reliant on brute force; their saliva contains a mix of toxic bacteria and proteins that can incapacitate prey, ensuring a successful hunt even if the initial attack does not immediately kill. Such documentaries provide compelling evidence of the Komodo dragon's adaptability and survival strategies, offering an in-depth view of their role in the ecosystem as both predators and scavengers. Through expert commentary and vivid footage, viewers gain a deeper understanding of these ancient reptiles, whose hunting prowess and dietary habits have earned them a fearsome reputation in the natural world.
2	Indirect positive(offering background)	Scientific studies describing the Komodo dragon's unique saliva, which contains venomous proteins that help subdue prey.	Scientific studies have delved into the fascinating biology of the Komodo dragon, revealing that its saliva is not just ordinary; it is laced with venomous proteins that play a critical role in its predatory success. This unique composition of the Komodo dragon's saliva is a subject of intrigue and has been analyzed to uncover the specific proteins responsible for its venomous properties. Research has identified several enzymes, including phospholipase A2, which contribute to the anticoagulant effects and cause rapid blood loss in prey. Additionally, these proteins are known to lower blood pressure, induce shock, and inhibit clotting, making it easier for the Komodo dragon to subdue its prey. Studies have shown that the venom's potency is comparable to that of venomous snakes, providing a significant advantage in the wild. This discovery has challenged previous assumptions that the Komodo dragon's hunting prowess was solely due to bacterial infection from its bite, highlighting the complexity and efficiency of its biological weaponry.
3	Negative Document	Studies on the dietary habits of other large reptiles, like crocodiles, without specific reference to Komodo dragons or their unique hunting strategies.	In examining the dietary habits of large reptiles, particularly crocodiles, we find fascinating insights into their feeding strategies and ecological roles. Crocodiles, being apex predators, exhibit a diet that is both diverse and opportunistic. Studies have shown that their diet primarily consists of fish, but they are also known to consume a wide range of prey including birds, mammals, and occasionally other reptiles. For example, research conducted in the Everglades reveals that American crocodiles (Crocodylus acutus) can adapt their diet based on availability, often preying on crustaceans and small mammals during specific seasons when these prey are more abundant. Furthermore, Nile crocodiles (Crocodylus niloticus) in Africa have been documented engaging in cooperative hunting strategies, which involve herding fish into shallow waters to facilitate easier capture. This behavior highlights their adaptability and intelligence in securing food resources. Unlike Komodo dragons, which employ a unique hunting strategy involving venom and ambush, crocodiles rely more on their stealth, powerful jaws, and aquatic prowess to capture prey. This ability to exploit a variety of food sources underscores their role as crucial components of their ecosystems, maintaining the balance by controlling the population of various species.

Table 9: Passage examples. The examples are generated by LIMRANK-SYNTHESIZER.

[Question]:
The question is about how to represent GDP by industry as a network graph, where sectors are connected by flows of money or goods, and seeks research or visualizations that illustrate inter-industry links, sources, drains, and economic loops.
[Passage Summary]:
The passage is about the extensive research on input–output models, including their extensions to non-linear relationships, inter-industry clustering, and macro-level economic analysis.
[Output]:
Okay, let's tackle this query. The user wants to know if the given passage is relevant to their question about viewing GDP as a network graph. The query mentions using industries like real estate and motor vehicles as examples of how sectors might be connected through transactions, creating a network.
Looking at the passage, it discusses the input-output model, which is a framework used to analyze economic transactions between industries. The passage talks about extending the model to non-linear relationships, the Hawkins-Simon condition, disaggregation into industry clusters, and empirical studies on coefficient identification. It also mentions extending the Leontief system to general equilibrium and decomposition methods.
Now, the user's main interest is in creating a network graph of GDP industries, showing connections between sectors. **The passage doesn't explicitly mention network graphs or visualization**. **However**, the input-output model is foundational for understanding how industries interact, which is exactly what the user is asking about. The passage's discussion of disaggregating industries and empirical work on coefficients relates to building a detailed network model. Even though the passage doesn't use the term n̈etwork graph,ẗhe underlying principles of the input-output model are relevant because they provide the structure and data needed to create such a visualization. The mention of general equilibrium and decomposition methods might also support analyzing the sources and drains in the network. Therefore, **even if the passage doesn't directly address network graphs**, it gives **foundational knowledge about the economic relationships** that would be necessary for creating one. So, the answer should be true.

Figure 11: Outputs from LimRank. The queries and the passages are given in summary format for clarity. The example shows that LimRank can distingush the passage with indirect correlation.

[Question]:
The question is about how airlines claiming to offer carbon-neutral flights explain and demonstrate these claims, specifically regarding their use of alternative fuels, transparency, and accounting methods to ensure that emissions are not simply shifted elsewhere.
[Passage]:
The passage is about the Qantas SAF Coalition Program, which supports the development of a domestic sustainable aviation fuel (SAF) industry in Australia by encouraging corporate contributions to offset the higher cost of SAF and reduce business travel emissions. **[Output]**:
Okay, let's tackle this. The user is asking whether the provided passage is relevant to the query about how airlines plan to make flights carbon-neutral. The query specifically mentions CNN's article from 2022 about Boeing's ecoDemonstrator jet and the airlines' vows to offer carbon-neutral flights. First, the passage talks about the SAF Coalition Program, which is about supporting the development of sustainable aviation fuel (SAF) in Australia. It mentions Qantas as part of the coalition, where corporations contribute to the cost of SAF, helping to make it more affordable. The program's goal is to reduce emissions by using SAF, which is a key method for carbon neutrality in aviation. Now, the query is asking how airlines explain making flights carbon-neutral. The passage directly discusses a program that involves using SAF to reduce emissions, which is a method mentioned in the query. Even though the passage doesn't explicitly say how airlines communicate their plans, it does provide information on a specific approach (SAF) that airlines might use. Since the query is about how they plan to achieve carbon neutrality, the mention of SAF as a method is relevant. The passage also talks about corporate contributions and making SAF affordable, which relates to the financial aspects of implementing such plans. However, the passage doesn't address transparency or accounting methods, which are part of the user's concern. But the question is whether the passage is relevant, not whether it answers all parts. The presence of SAF as a method is directly related to the query's focus on how airlines plan to make flights carbon-neutral. Therefore, the passage is relevant because it provides information on a specific strategy (SAF) that airlines might use, even if it doesn't detail the communication aspects.

Figure 12: Outputs from LimRank. The queries and the passages are given in summary format for clarity. The example shows that LimRank can also process the passage with direct correlation.