
Proceedings Paper:
Laksito, A., Alqarni, A. and Stevenson, M. (2026) Rank-ICL: Ranking-based In-context Learning for Search Result Explanation. In: Moffat, A., Scholer, F., Bi, K. and Clarke, C.L.A., (eds.) ICTIR '26: Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR). ICTIR '26: International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR), 25 Jul 2026, Melbourne, VIC, Australia. Association for Computing Machinery (ACM), pp. 199-205. ISBN: 9798400726002.

https://doi.org/10.1145/3805713.3820420

Rank-ICL: Ranking-based In-context Learning for Search Result Explanation

Arif Laksito
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
alaksito1@sheffield.ac.uk

Aali Alqarni
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
aaalqarni1@sheffield.ac.uk

Mark Stevenson
School of Computer Science
University of Sheffield
Sheffield, United Kingdom
mark.stevenson@sheffield.ac.uk

Abstract

Explanations in search results typically consist of text snippets or short passages presented alongside retrieved documents to help users efficiently assess relevance. While large language models (LLMs) have demonstrated strong performance across a wide range of language understanding and generation tasks, prior work on their use to generate explanations in search results remains relatively sparse. In this study, we investigate the use of decoder-only LLMs to generate explanations in the search results. To improve explanation quality in low-supervision settings, we introduce a ranking-based strategy for selecting informative few-shot examples in in-context learning. Rather than relying on randomly chosen demonstrations, relevant examples are dynamically retrieved based on retrieval functions to the input query–document pair. Evaluation on WikiSA and ExaRank shows that ranking-based few-shot prompting generally improves over zero-shot prompting and achieves competitive performance against random-shot prompting. However, its effectiveness varies across datasets, indicating that retrieval-based demonstration selection is beneficial but not uniformly superior in all settings.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; • **Computing methodologies** → *Natural language generation.*

Keywords

Explainable Information Retrieval, Large Language Models, Ranking Models, In-Context Learning

ACM Reference Format:
Arif Laksito, Aali Alqarni, and Mark Stevenson. 2026. *Rank-ICL*: Ranking-based In-context Learning for Search Result Explanation. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '26), July 25, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3805713.3820420

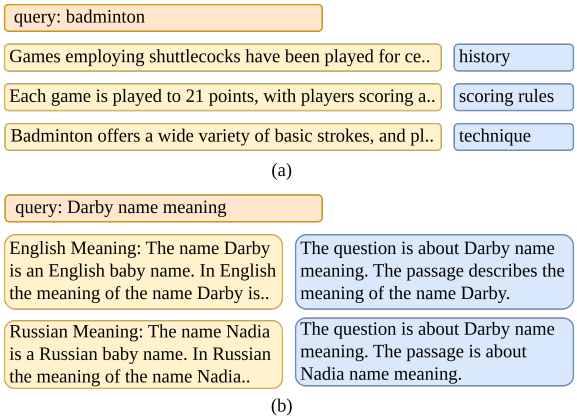

Figure 1: Examples of search result explanation generation. (a) Aspect-based explanations that summarise specific aspects of retrieved documents relevant to the query. (b) Relevance-based explanations that provide longer justifications describing overall document relevance.

1 Introduction

In search systems, users frequently submit under-specified queries with multiple potential interpretations of the user's intent [11, 15, 21]. This ambiguity often results in a diverse range of search results, requiring users to sift through numerous documents to find relevant information. Snippets were introduced to help users quickly assess the relevance of a document to their query [25]. These typically include the document's title, URL, and a brief summary of its contents, usually consisting of two to three lines. A study indicated that while snippets can enhance user interaction with search systems they often fall short of clearly explaining the relevance of the query to the retrieved documents [18]. Early work on explaining query–document relevance primarily focused on feature-based, such as term matches of query and document [17, 23, 28]. In contrast, abstractive or free-text explanations offer a more expressive alternative by generating natural language text that can articulate relevance beyond explicit feature attributions [1]. This setting is also related to rationale generation in NLP, where models produce natural-language justifications or supporting explanations for predictions [4, 16, 29], although our focus is specifically on query–document relevance in search results.

Free-text explanations style can vary in length, ranging from concise to more detailed descriptions. **Aspect-based explanations** generally consist of short phrases or keywords that capture specific aspects of a document relevant to the query, typically limited to one to five words [18, 31]. In contrast, **relevance-based explanations** are longer and more comprehensive. These explanations clarify both the query's intent and how the document addresses it, offering

ICTIR '26, Melbourne, VIC, Australia.

ACM ISBN 979-8-4007-2600-2/2026/07
https://doi.org/10.1145/3805713.3820420

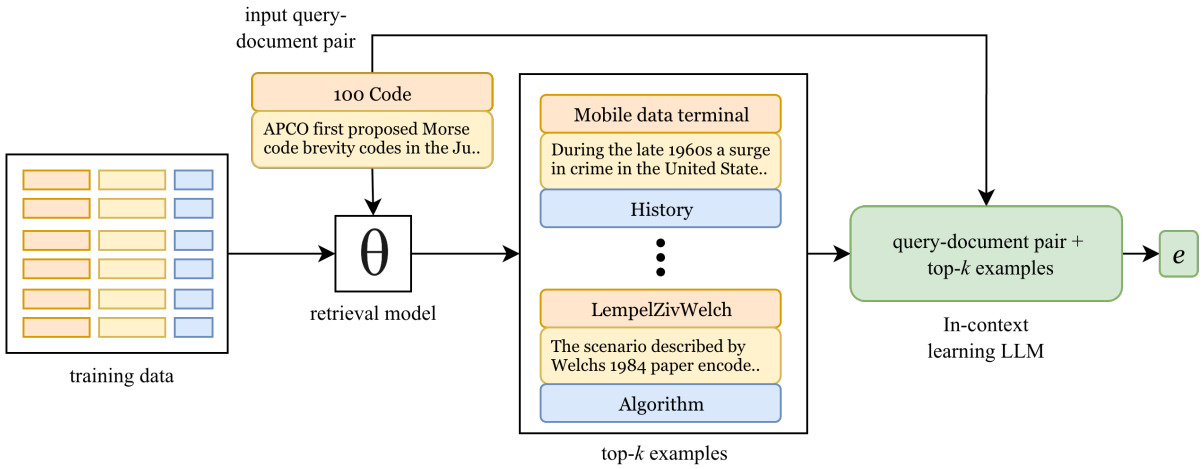

Figure 2: Rank-ICL framework overview. Given an input query–document pair, a retrieval component first produces the top-k similar query–document–explanation examples from a training corpus. These retrieved examples are incorporated into the prompt as in-context demonstrations, together with the target input. The LLM then generates an explanation by conditioning on both the input pair and the retrieved samples, without updating model parameters.

a broader and deeper explanation of the relevance [9]. In this work, we treat aspect-based and relevance-based explanations as two related but distinct search explanation settings, as illustrated in Figure 1. We focus on explanations that help users understand why a document is relevant to a query, rather than explaining the internal ranking mechanism.

Pre-trained large language models (LLMs), particularly those based on the transformer architecture [27], have shown strong performance across a wide range of text generation tasks. Recent decoder-only models, such as GPT [3] and LLaMA [26], generate text by predicting the next token in a sequence through large-scale pre-training. This mechanism enables LLMs to produce fluent and contextually coherent text conditioned on an input prompt. As a result, they have been widely applied to natural language generation tasks, including summarisation, question answering, dialogue generation, and explanation generation. Their ability to follow natural language instructions also makes them useful in low-supervision settings, where task-specific training data may be limited or costly to obtain.

In-context learning (ICL), also known as few-shot prompting, offers an alternative inference-time strategy by enabling task adaptation through conditioning on a small set of labelled examples provided in the input prompt, without updating model parameters [24]. Formally, given a test instance representation, x, frozen model parameters, ϕ, and a set of k labelled examples, $N_k(x;\theta)$, selected using a similarity function, θ, the prediction can be expressed as:

$$P(y \mid x;\phi,\theta) = f(x, N_k(x;\theta);\phi), \tag{1}$$

where $f(\cdot)$ denotes the model's inference function [13]. The choice of retrieval function, θ, whether lexical (e.g., tf–idf) or semantic, plays a critical role in determining ICL effectiveness.

Previous work has focused on generating aspect-based explanations by modifying transformer architectures and training them on large Wikipedia corpora [18]. Following this, Yu et al. [31] used a fine-tuned BART pre-trained model with a proposed architecture to improve performance. However, updating model parameters is computationally expensive, especially for LLMs with billions of parameters.

Recent work has shown that ICL can benefit from retrieval-based example selection. For example, Kapuriya et al. [13] use similarity-based retrieval and diversity-aware reranking to improve ICL on classification-style NLP tasks such as entailment, sentiment classification, and question classification, while Das et al. [6] retrieve similar problem-solution pairs to support ICL for code-quality estimation. However, these approaches focus on prediction or ranking tasks, where the model outputs a label, score, or ordering, rather than generating natural-language explanations. In contrast, we propose *Rank-ICL*, a ranking-based in-context learning framework for search result explanation generation, where in-context examples are selected through retrieval model. Furthermore, we provide a systematic analysis showing that the choice of retrieval function for example selection (e.g., lexical vs. semantic) has an impact on explanation quality and behaviour across different LLMs. This highlights demonstration selection as a critical yet previously under-examined component in LLM-based explanation generation for search.

2 Rank-ICL

Task Formulation. Let $\mathcal{Q}$ denote the space of queries and $\mathcal{D}$ the space of documents. Given a query $q \in \mathcal{Q}$ and a retrieved document $d \in \mathcal{D}$, the goal of search result explanation generation is to produce an explanation e that articulates the relevance of d with respect to q.

We model explanation generation as a conditional text generation problem, where an LLM estimates the probability distribution

$$P(e \mid q, d), \tag{2}$$

and generates an explanation by conditioning on the query–document pair.

Under the ICL setting, the model is conditioned on a small set of examples, $N_k(q, d; \theta) = \{(q_i, d_i, e_i)\}_{i=1}^{k}$, each of which consists of a query, document and explanation. The set of examples is selected from a training corpus based on similarity to the input query–document pair. The generation process can be expressed as:

$$P(e \mid q, d; \phi, \theta) = f(q, d, N_k(q, d; \theta); \phi), \tag{3}$$

where ϕ denotes the frozen parameters of the pre-trained language model.

The explanation, e, may take different forms depending on the task setting. In the aspect-based setting, e is a short textual description highlighting specific aspects of the document relevant to the query. In the relevance-based setting, e is a longer justification explaining the overall relevance or non-relevance of the document. An overview of the Rank-ICL framework is shown in Figure 2.

3 Experiments

Our experiments are designed to evaluate the effectiveness of Rank-ICL for search result explanation generation. Specifically, we investigate (i) whether Rank-ICL improves explanation quality compared to zero-shot and random few-shot prompting; (ii) which retrieval models are most effective for selecting informative in-context examples; and (iii) how different LLMs affect the quality of the generated explanations.

Datasets. We conduct experiments on two benchmark datasets for search result explanation generation:
WikiSA [18] is used for aspect-based explanation generation. It is derived from the March 2024 English Wikipedia dump, where article titles are treated as queries and sections are treated as documents. Section headings serve as aspect-based explanations. Due to the lack of publicly available source data, we reconstruct the dataset following the original methodology. The reconstructed dataset contains 40.8k training instances, 4.6k validation instances, and 9.9k test instances.
ExaRank [9] is used for relevance-based explanation generation. It is based on the MS MARCO passage ranking dataset, with explanations automatically generated using a frontier large language model, `text-davinci-002`, with few-shot prompting. The original study used these explanations to improve ranking performance, rather than directly evaluating them as generated explanations. The dataset contains 21.6k training instances, 2.4k validation instances, and 6k test instances.
Due to the computational cost of LLM inference, we evaluate all methods on a fixed subset of 2k validation instances and 2k test instances for each dataset. The same evaluation subset is used across all models and experimental settings to ensure a fair comparison.

Baselines. Performance was compared against the following previous approaches to explanation generation:
Transformer [27]: The original encoder–decoder Transformer architecture for sequence-to-sequence generation. The model is initialised from scratch (random initialisation) and trained on the respective training splits of each dataset. Input text is tokenised using the BERT tokenizer to ensure consistency with prior work.
GenEx [18]: An encoder–decoder model that incorporates a query attention mechanism and query-masked decoding to generate query-focused explanations. The original architecture is used [18].
LiEGe [31]: A listwise explanation generation model that jointly explains documents in a ranked list by modelling cross-document interactions and multi-granularity representations. The model is initialised using pre-trained BERT or BART checkpoints, following the original implementation. We limit this implementation to WikiSA, as the ExaRank dataset provides single relevance-based explanations without explicit aspect decomposition, making it unsuitable for listwise multi-aspect modelling.
All baseline models are trained using tuned hyper-parameters, including learning rate, number of epochs, warm-up steps, and dropout rate. Model selection is performed based on validation performance, and early stopping is applied to prevent over-fitting. This ensures that all supervised baselines are reasonably optimised for a fair comparison.

Table 1: Best decoding settings for each LLM on the WikiSA and ExaRank datasets. These settings were used consistently for zero-shot, random-shot, and Rank-ICL prompting.

LLMs	WikiSA	ExaRank
LLaMA	temp 0.2, top-p 0.9, top-k 10, penalty 1.1	temp 0.1, top-p 0.8, top-k 0, penalty 1.0
Qwen	temp 0.2, top-p 0.9, top-k 10, penalty 1.0	temp 0.1, top-p 0.8, top-k 10, penalty 1.0
Mistral	temp 0.1, top-p 0.8, top-k 10, penalty 1.0	temp 0.1, top-p 0.8, top-k 10, penalty 1.0

Retrieval Models. We consider three retrieval models to select in-context examples. As lexical baselines, we use TF–IDF and BM25 [20] indices constructed using a BM25 library.[1] For semantic retrieval, the Sentence-BERT (SBERT) [19] neural ranker is used to encode texts into dense vector representations and rank candidate examples based on semantic similarity. The `all_datasets_v3_mpnet-base` model[2] is used for embedding generation. This model has been shown to produce high-quality sentence embeddings useful for semantic similarity and retrieval tasks [5]. Nearest-neighbour search using FAISS [8] is applied to enable efficient similarity search over dense vectors.

LLM Details. We use instruction-tuned LLMs from three widely adopted open-source model families: LLaMA3.1-8B [10], Qwen2-7B [30], and Mistral-7B-v0.3 [12]. These models represent strong and commonly used baselines for instruction-following generation, and differ in architecture design and pre-training data, while

[1] https://huggingface.co/blog/xhluca/bm25s
[2] https://huggingface.co/flax-sentence-embeddings/all_datasets_v3_mpnet-base

Table 2: Comparison of baselines, zero-shot, random-shot, and Rank-ICL prompting strategies across two distinct search explanation settings: WikiSA for aspect-based explanation generation and ExaRank for relevance-based explanation generation. Bold values indicate the best performance for each LLM under the corresponding dataset. Values marked with † denote the best overall score for each metric within the corresponding dataset. Results marked with * indicate statistically significant improvements over the Transformer baseline ($p < 0.05$, paired t-test with Bonferroni correction).

		WikiSA			ExaRank		
Models	Method	ROUGE-1	BERTScore	METEOR	ROUGE-1	BERTScore	METEOR
Transformer [27]	Fine-tuned	.2341	.6339†	.1497	.5367	.6325	.4530
GenEx [18]	Fine-tuned	.1828*	.6072	.1118*	.5214	.6102*	.4274*
LiEGe (BERT) [31]	Fine-tuned	.2244	.4049*	.1352*	-	-	-
LiEGe (BART) [31]	Fine-tuned	.2503*	.4799*	.2212*†	-	-	-
LLaMA	Zero-shot	.0481*	.4405*	.0358*	.3136*	.6013*	.3848*
	Random-shot	.1620*	.4879*	.1190*	**.4489***	**.6735***	**.4979***
	Rank-ICL(SBERT)	.1779*	.5025*	.1190*	.4221*	.6501*	.4574
	Rank-ICL(BM25)	**.1909***	**.5074***	**.1286***	.4226*	.6499*	.4576
	Rank-ICL(TF–IDF)	.1774*	.4987*	.1165*	.4229*	.6512*	.4592
Qwen	Zero-shot	.1600*	.5238*	.1554*	.3463*	.6363	.4272*
	Random-shot	.1727*	.5235*	.1603*	**.6610*†**	**.7919*†**	**.6955*†**
	Rank-ICL(SBERT)	.2656*	.5984*	.2010*	.6033*	.7646*	.6382*
	Rank-ICL(BM25)	**.2940*†**	**.6106***	**.2200***	.5965*	.7614*	.6319*
	Rank-ICL(TF–IDF)	.2639*	.5955*	.2021*	.5961*	.7620*	.6293*
Mistral	Zero-shot	.0592*	.4234*	.0729*	.3467*	.6287	.4153*
	Random-shot	.1768*	.4842*	.1441	.6065*	.7683*	.6371*
	Rank-ICL(SBERT)	.2320*	.5048*	.1726*	**.6123***	**.7718***	**.6514***
	Rank-ICL(BM25)	**.2480***	**.5118***	**.1781***	.6028*	.7662*	.6405*
	Rank-ICL(TF–IDF)	.2340	.5037*	.1684*	.5982*	.7641*	.6370*

operating at a comparable parameter scale (7B–8B). All models are deployed with 4-bit quantised weights for computational efficiency [7].

Prompting Strategy. We begin with zero-shot prompting to establish a baseline decoding configuration. Decoding hyper-parameters are tuned on a validation set for each dataset by exploring both deterministic and stochastic generation settings. Specifically, we consider deterministic decoding (`do_sample = false`), as well as stochastic decoding (`do_sample = true`) with temperature $\in \{0.1, 0.2, 0.3\}$, top-$p \in \{0.8, 0.9\}$, top-$k \in \{0, 10, 20\}$, and repetition-penalty $\in \{1.0, 1.1\}$. The best-performing configuration is selected based on validation performance for each dataset. To ensure a fair comparison across prompting strategies, the selected decoding configuration is fixed and consistently applied to random-shot and Rank-ICL prompting for all models. We use $k = 3$ in-context examples for both random-shot and Rank-ICL prompting, following prior work showing that this provides a good balance between demonstration diversity and performance [22]. Table 1 shows the best decoding settings for each model on both datasets, which were used consistently across all prompting strategies. Our implementation is publicly available to facilitate reproducibility and further research.[3]

Evaluation Metrics. Following prior work on explanation generation [18, 31], we evaluate explanation quality using ROUGE-1, METEOR, and BERTScore. ROUGE-1 measures unigram overlap between generated and reference explanations, capturing lexical similarity and content coverage [14]. METEOR extends n-gram matching by incorporating synonymy, stemming, and paraphrase information, making it more sensitive to linguistic variation [2]. BERTScore computes semantic similarity using contextual embeddings from pre-trained language models, capturing meaning beyond surface form [32]. Together, these metrics assess both lexical and semantic aspects of explanation quality.

[3]https://github.com/ariflaksito/rank-icl

4 Results and Discussions

Table 2 summarises the performance of Rank-ICL across different retrieval models and LLMs for search result explanation generation.

Rank-ICL compared to Baselines, Zero-shot and Random few-shot Prompting. Few-shot prompting consistently outperforms zero-shot prompting across all evaluation metrics, LLMs, and datasets. These findings indicate that conditioning LLMs on in-context examples is beneficial for explanation generation, regardless of whether the demonstrations are selected randomly or through retrieval-based ranking. Rank-ICL provides the clearest gains over random-shot prompting on WikiSA, where ranking-based selection of demonstrations generally leads to higher performance. This suggests that retrieved examples provide more informative and structured context for aspect-based explanation generation, where lexical or semantic similarity between query–document pairs can help guide the model toward more specific aspect phrases.

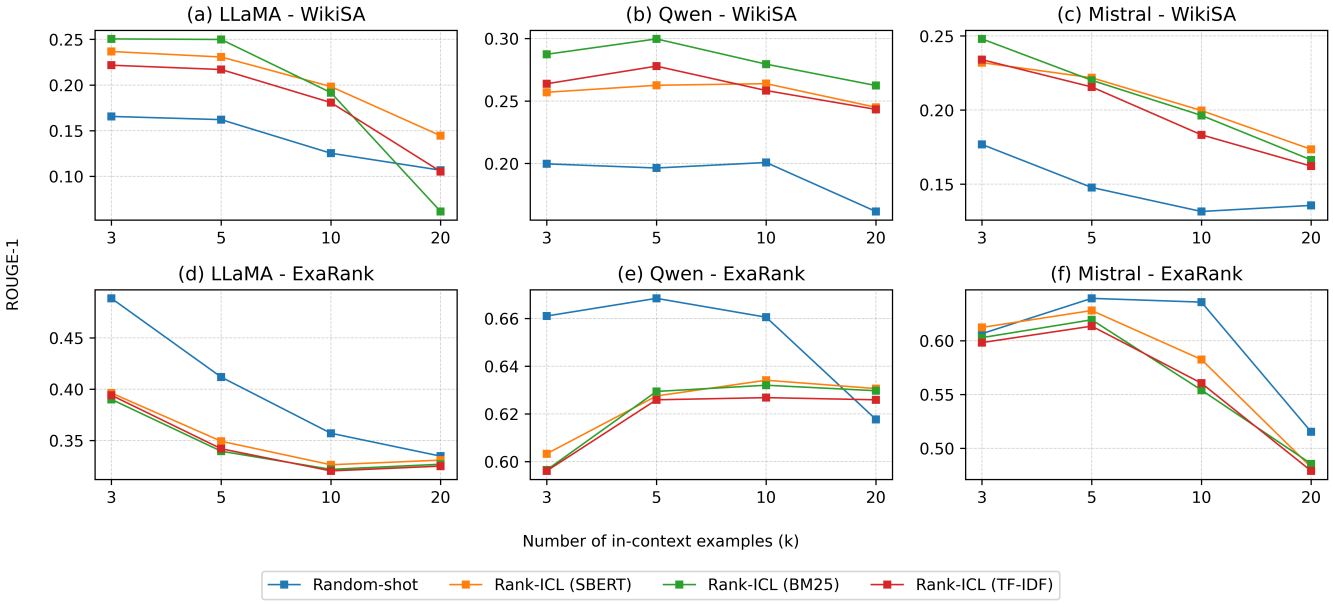

Figure 3: Effect of the number of in-context examples on ROUGE-1 for Random-shot and Rank-ICL with different retrieval models across LLMs and datasets.

However, the advantage of Rank-ICL is less consistent on ExaRank. For some LLMs, notably Qwen and LLaMA, random-shot prompting outperforms Rank-ICL. This indicates that when the underlying model is sufficiently strong, or when the dataset contains relatively consistent explanation patterns, randomly sampled demonstrations may already provide adequate contextual signals. These results highlight that the effectiveness of Rank-ICL depends on both the dataset characteristics and the underlying LLM.

In comparison to supervised baselines, LiEGe (BART) achieves the strongest performance on METEOR, while the Transformer model attains the highest BERTScore on the WikiSA dataset. This is expected, as supervised baselines are trained end-to-end and optimises all model parameters for the explanation generation task, enabling it to capture aspect relationships across documents. However, this comes at a significant computational cost, requiring extensive training and hyper-parameter tuning to reach optimal performance.

In contrast, Rank-ICL operates without training and does not require updating model parameters. Despite this, it achieves competitive performance relative to supervised approaches, demonstrating that ranking-based in-context learning can effectively guide LLMs for explanation generation without full model training. This highlights a trade-off between performance and efficiency, particularly in low-resource environments.

Impact of Retrieval Models. The effectiveness of retrieval models varies across the two dataset settings. On WikiSA, BM25 consistently achieves the best performance across LLMs, suggesting that lexical retrieval is well suited for selecting demonstrations for aspect-based explanations. This is likely because WikiSA explanations are short aspect phrases, where exact or near-exact term overlap between query–document pairs can provide useful guidance.

On ExaRank, the pattern is less uniform. Semantic retrieval with SBERT achieves the best retrieval performance for Mistral, indicating that semantic similarity can be particularly useful for relevance-based explanations, where the model needs to generate a broader justification rather than a short aspect phrase. However, this trend is not consistent across all LLMs, as BM25 and TF–IDF remain competitive in several settings. Overall, no single retrieval model dominates across all datasets and models.

Effect of Different LLMs. Performance differences are also observed across LLMs. Overall, Qwen achieves the strongest performance among the evaluated models, followed by Mistral and LLaMA. On WikiSA, Qwen obtains the best results across most metrics, indicating that it is more effective for generating concise aspect-based explanations. A similar pattern is observed on ExaRank, where Qwen also achieves the highest overall scores, particularly under few-shot prompting. Mistral generally performs competitively and often remains close to Qwen, while LLaMA tends to obtain lower scores across both datasets.

These results indicate that the underlying LLM has a clear impact on explanation quality. Although all models benefit from in-context examples compared with zero-shot prompting, stronger instruction-tuned models appear to generate more accurate and better-aligned explanations.

Effect of Different Number of Examples. We further analyse the effect of varying the number of in-context examples on explanation quality for both random few-shot prompting and Rank-ICL. Results are shown in Figure 3. They demonstrate that increasing k does

not yield consistent performance gains, with the best results often observed at smaller values (e.g., $k = 3$ or 5). Performance frequently plateaus or declines as k increases, suggesting that additional examples may introduce less relevant context. It appears that increasing k from 3 to 5 for the Qwen model leads to performance improvement on two datasets, but using more than 5 examples results in either stable or degraded performance for Rank-ICL. Overall, these findings indicate that a small number of well-ranked examples is generally sufficient, and that example quality is more important than quantity.

Limitations. This study has several limitations. First, Rank-ICL does not uniformly outperform random-shot prompting, particularly on ExaRank for some LLMs, indicating that retrieval-based example selection is most beneficial when retrieved demonstrations provide additional task-specific signals beyond those available from random examples. Second, our experiments focus on 7B–8B open-source LLMs, and future work should examine whether the observed retrieval-function trends remain stable for larger models.

5 Conclusion

Overall, the results demonstrate that Rank-ICL is an effective framework for search result explanation generation. Ranking-based selection of in-context examples consistently improves performance over zero-shot and achieves competitive performance compared to random-shot prompting, particularly for aspect-based explanations. Our analysis shows that the effectiveness of retrieval strategies depends on the explanation type: BM25 performs best for aspect-based explanations, while semantic retrieval (SBERT) is more beneficial for relevance-based explanations. We also find that model choice remains important, with Qwen achieves the best performance followed by Mistral and LLaMA across datasets. In addition, increasing the number of in-context examples does not always improve performance. A small number of high-quality examples is typically sufficient, highlighting that example quality is more critical than quantity. Compared to supervised approaches such as LiEGe and Transformer, which achieve strong performance through full model training, Rank-ICL provides a training-free alternative with a favourable trade-off between effectiveness and computational cost.

Acknowledgments

We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing. The first author also acknowledges the support provided by Indonesian Education Scholarship, Center for Higher Education Funding and Assessment and Indonesia Endowment Fund for Education through a doctoral studentship.

References

[1] Avishek Anand, Lijun Lyu, Maximilian Idahl, Yumeng Wang, Jonas Wallat, and Zijian Zhang. 2022. Explainable Information Retrieval: A Survey. (2022). arXiv: 2211.02405v1.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909/

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[4] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/hash/4c7a167bb329bd92580a99ce422d6fa6-Abstract.html

[5] Boqi Chen, Kua Chen, Yujing Yang, Afshin Amini, Bharat Saxena, Cecilia Chávez-García, Majid Babaei, Amir Feizpour, and Dániel Varró. 2023. Towards Improving the Explainability of Text-based Information Retrieval with Knowledge Graphs. doi:10.48550/arXiv.2301.06974 arXiv:2301.06974 [cs].

[6] Susmita Das, Madhusudan Ghosh, Priyanka Swami, Debasis Ganguly, and Gul Calikli. 2025. In-Context Learning as an Effective Estimator of Functional Correctness of LLM-Generated Code. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Padua Italy, 2890–2894. doi:10.1145/3726302.3730212

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems* 36 (Dec. 2023), 10088–10115. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html

[8] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss library. doi:10.48550/arXiv.2401.08281 arXiv:2401.08281 [cs].

[9] Fernando Ferraretto, Thiago Laitz, Roberto Lotufo, and Rodrigo Nogueira. 2023. ExaRanker: Synthetic Explanations Improve Neural Rankers. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2409–2414. doi:10.1145/3539618.3592067

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, and Korenev. 2024. The Llama 3 Herd of Models. doi:10.48550/arXiv.2407.21783 arXiv:2407.21783 [cs].

[11] Mayu Iwata, Tetsuya Sakai, Takehiro Yamamoto, Yu Chen, Yi Liu, Ji-Rong Wen, and Shojiro Nishio. 2012. AspecTiles: tile-based visualization of diversified web search results. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, Portland Oregon USA, 85–94. doi:10.1145/2348283.2348298

[12] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. doi:10.48550/arXiv.2310.06825 arXiv:2310.06825 [cs].

[13] Janak Kapuriya, Manit Kaushik, Debasis Ganguly, and Sumit Bhatia. 2025. Exploring the Role of Diversity in Example Selection for In-Context Learning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Padua Italy, 2962–2966. doi:10.1145/3726302.3730194

[14] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[15] Sean MacAvaney, Craig Macdonald, Roderick Murray-Smith, and Iadh Ounis. 2021. IntenT5: Search Result Diversification using Causal Language Models. http://arxiv.org/abs/2108.04026 arXiv:2108.04026 [cs].

[16] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. doi:10.48550/arXiv.2004.14546 arXiv:2004.14546 [cs.CL].

[17] Sayantan Polley, Atin Janki, Marcus Thiel, Juliane Hoebel-Mueller, and Andreas Nuernberger. 2021. ExDocS: Evidence based Explainable Document Search. https://csr21.github.io/polley-csr2021.pdf

[18] Razieh Rahimi, Youngwoo Kim, Hamed Zamani, and James Allan. 2021. Explaining Documents' Relevance to Search Queries. (Nov. 2021). https://arxiv.org/abs/2111.01314v1 arXiv: 2111.01314.

[19] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. doi:10.48550/arXiv.1908.10084 arXiv:1908.10084 [cs].

[20] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389. doi:10.1561/1500000019

[21] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search Result Diversification. *Found. Trends Inf. Retr.* 9, 1 (March 2015), 1–90. doi:10.1561/1500000040

[22] Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed Fact Checking using RAG and Few-Shot In-Context Learning with LLMs. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 91–98. doi:10.18653/v1/2024.fever-1.10

[23] Jaspreet Singh and Avishek Anand. 2019. EXS: Explainable search using local model agnostic interpretability. *WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining* (Jan. 2019), 770–773. doi:10.1145/3289600.3290620 arXiv: 1809.03857 ISBN: 9781450359405.

[24] Nilanjan Sinhababu, Andrew Parry, Debasis Ganguly, Debasis Samanta, and Pabitra Mitra. 2024. Few-shot Prompting for Pairwise Ranking: An Effective Non-Parametric Retrieval Model. doi:10.48550/arXiv.2409.17745 arXiv:2409.17745 [cs].

[25] Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, Melbourne Australia, 2–10. doi:10.1145/290941.290947

[26] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. doi:10.48550/arXiv.2307.09288 arXiv:2307.09288 [cs].

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[28] Manisha Verma and Debasis Ganguly. 2019. LIRME: Locally Interpretable Ranking Model Explanation. (July 2019), 1281–1284. doi:10.1145/3331184.3331377

[29] Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring Association Between Labels and Free-Text Rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10266–10284. doi:10.18653/v1/2021.emnlp-main.804

[30] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671* (2024).

[31] Puxuan Yu, Razieh Rahimi, and James Allan. 2022. Towards Explainable Search Results: A Listwise Explanation Generator. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022). doi:10.1145/3477495 Place: New York, NY, USA ISBN: 9781450387323.

[32] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. doi:10.48550/arXiv.1904.09675 arXiv:1904.09675 [cs].