An Investigation of Prompt Variations for Zero-shot LLM-based Rankers

Shuoqi Sun[1] , Shengyao Zhuang[2], Shuai Wang[3], Guido Zuccon[3]

1 RMIT University, Melbourne, Australia
shuoqi.sun@student.rmit.edu.au

2 CSIRO, Australia
shengyao.zhuang@csiro.au

3 The University of Queensland, St. Lucia, Australia
{shuai.wang, g.zuccon}@uq.edu.au

Year :- 2025
Number of Citations :- 5

Introduction - LLM Based Zero Shot Rankers

- Investigates the use of LLMs to create zero shot rankers [1-3].
- Zero shot rankers means methods which does not involve any specific fine tuning or training for ranking.
- These rankers operate by following the instructions in the prompt which include the query and the k documents that should be considered for ranking.
- The rankers use the LLM to generate answer that comply with the given instruction.
- Finally the generated answer contains the ranking provided by the method or the logits of the answers are used to infer the ranking.

[1] Ma, X., Zhang, X., Pradeep, R., Lin, J.: Zero-shot listwise document reranking with a large language model. arXiv preprint arXiv:2305.02156 (2023).
[2] Sun, W., Yan, L., Ma, X., Ren, P., Yin, D., Ren, Z.: Is chatgpt good at search? investigating large language models as re-ranking agent. arXiv preprint arXiv:2304.09542 (2023).
[3] Tang, R., Zhang, X., Ma, X., Lin, J., Ture, F.: Found in the middle: Permutation self-consistency improves listwise ranking in large language models. arXiv preprint arXiv:2310.07712 (2023).

Introduction - Different types of Zero shot rankers

- Four families of zero shot rankers. They differ in the ranking algorithm implemented in the instructions given in the prompt.
  - Pointwise [4].
  - Pairwise [5].
  - Listwise [6].
  - Setwise [7].
- Within each family, one or more methods are proposed differing in the backbone LLM and wording of the prompt.
- Recent works have shown that fixing on how the ranking is obtained (generation vs logits) and backbone LLM, setwise is the most effective. Depending on the dataset listwise or pairwise are also effective, with pointwise providing lower effectiveness overall.

[4] Zhuang, H., Qin, Z., Hui, K., Wu, J., Yan, L., Wang, X., Bendersky, M.: Beyond yes and no: Improving zero-shot pointwise llm rankers via scoring fine-grained relevance labels. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2024).
[5] Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., et al.: Large language models are effective text rankers with pairwise ranking prompting. arXiv preprint arXiv:2306.17563 (2023).
[6] Ma, X., Zhang, X., Pradeep, R., Lin, J.: Zero-shot listwise document reranking with a large language model. arXiv preprint arXiv:2305.02156 (2023).
[7] Zhuang, S., Zhuang, H., Koopman, B., Zuccon, G.: A setwise approach for effective and highly efficient zero-shot ranking with large language models. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 38–47 (2024).

Introduction

- The previous works did recognise that that the use of different backbone LLMs in previous work biased the comparison between methods, they did not identify that the actual prompts used by the different rankers differed not just in terms of the words used to describe the ranking algorithm. For example, the difference in “role-playing” component.
- What is the effect of such differences in wording used in the prompts?
- Are differences in effectiveness due to the actual ranking algorithm, or they are due to the choice of words used in the prompts?
- Are differences due to LLM characteristics such as backbone and size?

Related Works

- Sensibility of LLMs to prompt formulation.
  - Dependency between prompt and effectiveness [8].
  - Effect of prompt variations in relevance labeling [9].
  - Relationships between prompt strategies and social biases in LLMs [10] .
  - Effect of the prompts in a range of NLP tasks [11].
- Prompt Optimisation and Self-optimisers [12].
  - Using LLMs to iteratively refine prompts.
  - This is beyond the scope of this.
- Zero shot LLM rankers.
  - Pointwise (Generation, likelihood) [4].
  - Pairwise (Relevance of 2 documents to the query) [5].
  - Listwise ranking [6].
  - Setwise ranking [7].

[4] huang, H., Qin, Z., Hui, K., Wu, J., Yan, L., Wang, X., Bendersky, M.: Beyond yes and no: Improving zero-shot pointwise llm rankers via scoring fine-grained relevance labels. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2024).
[5] Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., et al.: Large language models are effective text rankers with pairwise ranking prompting. arXiv preprint arXiv:2306.17563 (2023).
[6] Ma, X., Zhang, X., Pradeep, R., Lin, J.: Zero-shot listwise document reranking with a large language model. arXiv preprint arXiv:2305.02156 (2023).
[7] Zhuang, S., Zhuang, H., Koopman, B., Zuccon, G.: A setwise approach for effective and highly efficient zero-shot ranking with large language models. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 38–47 (2024).
[8] Kim, J., Park, S., Jeong, K., Lee, S., Han, S.H., Lee, J., Kang, P.: Which is better? exploring prompting strategy for llm-based metrics. arXiv preprint arXiv:2311.03754 (2023)
[9] Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1930–1940 (2024).
[10] Kamruzzaman, M., Kim, G.L.: Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. arXiv preprint arXiv:2404.17218 (2024)
[11] Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., Stanovsky, G.: State of what art? a call for multi-prompt llm evaluation. Transactions of the Association for Computational Linguistics 12, 933–949 (2024)
[12] Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q.V., Zhou, D., Chen, X.: Large language models as optimizers. arXiv preprint arXiv:2309.03409 (2023).

Methodology - Taxonomy of Ranking prompt components

1) Evidence (EV) - these are the query and the associated passages to rank.
2) Task Instruction (TI) - the instructions associated to the specific ranking strategy: these outline to the LLM the algorithmic steps to follow to produce a ranking. Example wordings include “which passage is more relevant” (pairwise) and “is this passage relevant to the query” (pointwise).
3) Output Type (OT): the instructions that specify the format of the output the LLM needs to generate. For example, for pointwise ranking the LLM could be instructed to generated a Yes/No or a True/False answer.
4) Tone Words (TW): words that express a positive, negative, or neutral connotation and that help express the attitude of the prompt author towards the ranking instruction, e.g., “please” or “you better get this right or you will be punished”.
5) Role Playing (RP): a description of the tool implemented by the LLM, used to make the LLM “impersonate” that role.

Methodology - Taxonomy of Ranking prompt components

- EV always present.
- TI and OT are ranker family dependent.
- TW and RP can be applied to any ranker family.

Prompt Variations

			Wording Alternatives				
Component	Ranker	None (0)	1	2	3	4	5
Task Instruction (TI)	pointwise	-	Does the passage answer the query?	Is this passage relevant to the query?	For the following query and document, judge whether they are relevant.	Judge the relevance between the query and the document.	-
	pairwise	-	Given a query, which of the following two passages is more relevant to the query?	-			
	listwise	-	Rank the {num} passages based on their relevance to the search query.	Sort the Passages by their relevance to the Query.	I will provide you with {num} passages, each indicated by number identifier []. Rank the passages based on their relevance to query.	-	
	setwise	-	Which one is the most relevant to the query.	-			
Output Type (OT)	pointwise		Judge whether they are "Highly Relevant", "Somewhat Relevant", or "Not Relevant".	From a scale of 0 to 4, judge the relevance.	Answer 'Yes' or 'No'.	Answer True/False.	-
	pairwise	-	Output Passage A or Passage B.	-			
	listwise	-	Sorted Passages = [	The passages should be listed in descending order using identifiers. The most relevant passages should be listed first. The output format should be [] >[], e.g., [1] >[2].	-		
	setwise	-	Output the passage label of the most relevant passage.	Generate the passage label.	Generate the passage label that is the most relevant to the query, then explain why you think this passage is the most relevant.	-	
Tone Words (TW)	All	✓	You better get this right or you will be punished.	Only output the ranking results, do not say any word or explanation.	Please	Only	Must
Role Playing (RP)	All	✓	You are RankGPT, an intelligent assistant that can rank passages based on their rele-	-			

Methodology - Different Approaches in Ordering

- Evidence Ordering (EO): the relative ordering of the query and passage(s) provided to the LLM – whether the query is given first, followed by the passage(s), which we label as QF, or vice versa, passage(s) followed by the query (labelled PF).
- Position of Evidence (PE): instruction to specify the position of the evidence in the prompt – at the beginning (B) or at the end of the prompt (E).

Prompt templates based on ordering and components

EO/PE	B	E
QF	RP+ TI (Q) + P + TW+ OT	RP+ TW+ OT + TI (Q) + P
PF	RP+ P + TI (Q) + TW+ OT	RP+ TW+ OT + P + TI (Q)

Building Prompt variations

- There are 4 templates according to the ordering.
- For each template, all possible instantiations are considered.
- In total there are 1248 prompt variations tried out.
  - 768 unique prompts for pointwise.
  - 48 for pairwise.
  - 288 for listwise.
  - 144 for setwise.
- Minor modifications were done to prompts from previous works. For example, enclosing query in quote.

Experimental Settings

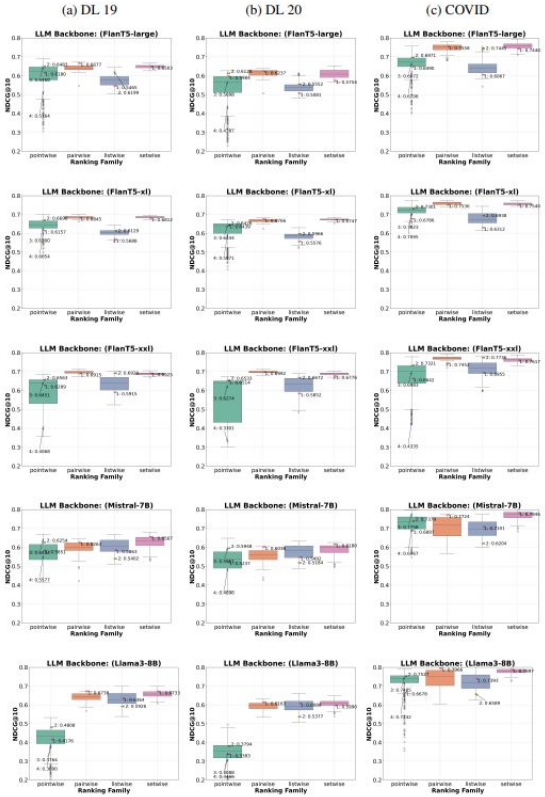
- 2 stage ranking pipeline.
  - BM25 implementation from Pyserini [13] to retrieve the top 100 documents for a query (k1=0.9, b=0.4).
  - LLM ranker to re rank those 100 documents.
- Instruction tuned checkpoints of open LLMs used as LLM backbone.
  - Flan-T5 [14].
  - Mistral 7B [15].
  - Llama3 8B [16].
- Also for Flan-T5, checkpoints of different sizes are considered. They are Large (783M), XL (2.85B) and XXL (11.3B).

[13] Lin, J., Ma, X., Lin, S.C., Yang, J.H., Pradeep, R., Nogueira, R.: Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2356–2362 (2021).
[14] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research 25(70), 1–53 (2024)
[15] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
[16] AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/ main/MODEL_CARD.md

Datasets

- 3 datasets are used to evaluate these prompts.
  - TREC Deep Learning (DL) 2019 (43 queries)
  - TREC Deep Learning (DL) 2019 (48 queries)
  - BeIR TREC COVID (50 queries)
- Analysis was done using nDCG@10, the primary metric across these datasets, and tested statistical significance with a paired, two-tails t-test.
- Flan-T5 models are constrained by a maximum input length of 512 tokens. So, for equitable comparison, query length was standardized to 20 words and document length to 80 words. This was decided by analysing the document lengths in the datasets.
- Even though this results in loss of information, as the primary goal is to have equitable comparison rather than actually ranking those original documents, it could be acceptable.

Result Analysis



Result Analysis

- Are there better prompts?
  - Prompts that can achieve higher effectiveness than the original prompts across all cases (with most differences being statistically significant), with the exception of listwise and pairwise on COVID when the Llama 3 backbone is used.
  - For the pairwise approach, only prompts for the FlanT5 - XXL model on the COVID dataset show significant differences, likely due to the lower effectiveness variation inherent in the pairwise method.
  - There are cases where the original prompt was the worst among those considered for the specific ranking family: for example for the listwise prompt evaluated on DL 19 and DL 20 with the FlanT5-Large backbone.

Result Analysis - Characteristics of best Prompts

- Task Instruction and Output Type
  - Pointwise ranking does not show a consistent optimal choice.
    - Task Instruction 2 showed better results.
    - Output type 3 is less optimal.
  - Listwise ranking shows that optimal choice varies by dataset and LLM backbone.
  - Setwise ranking shows that Output Type #3 is the most common output type among top performing prompts appearing in 53 % cases.
- Tone Words
  - A uniform influence of tone words on prompt effectiveness was observed.
  - No consistent patterns the influence of LLM backbones or datasets on the effectiveness related to tone words.
  - Including a tone word in the prompt led to increased effectiveness in 82% of the cases.

Result Analysis - Characteristics of best Prompts

- Role playing.
  - Role playing leads to best effectiveness for pointwise (80 %) and pairwise (66 %) prompts.
  - Mixed effects on set-wise.
  - Not associated with best effectiveness in listwise (13 %).
  - 55% of the prompts with highest effectiveness include role playing wording.
- Evidence ordering.
  - For pointwise ranking, presenting passage text before query text is preferred in 86% of top-performing prompts.
  - For other types of rankers, the difference is not clear.
  - Considering model backbones, Flan-T5 tends to perform best when presented with passage text before query text (66% of cases).

Result Analysis - Characteristics of best Prompts

- Position of Evidence
  - Among the best prompts, there tend to be an overall preference for prompts that provide the evidence at the beginning (before any other instruction): this is the case in 63% of the best prompts.
  - Pointwise and listwise prompts exhibiting more often this pattern (73% and 67% respectively).
  - Across all datasets, most best prompts for the FlanT5-XXL backbone have evidence at the beginning.
  - This is also the for Llama3-8B for DL19 and COVID and for Mistral-7B for COVID.

Result Analysis

- Which ranking method is most effective?
  - The best performing rankers were set-wise and pairwise (depending on dataset and backbone), followed by listwise and then pointwise, which were distinguishably worse.
  - However, experiments show that pointwise can be as competitive as these previous methods if instructed with specific prompts and this is the case in all datasets and backbones with the exception of Llama 3.
  - For Llama 3 on DL datasets, we observe that pointwise significantly underperforms other methods.

Result Analysis

- Are ranking methods stable?
  - Pointwise methods display the largest variability in effectiveness due to prompt variations, with some prompt variations delivering poor effectiveness.
  - Set-wise and pairwise do better showing lower variability.
  - Set-wise shows low variability across all datasets and across all LLM backbones.
  - Pairwise shows higher variability in specific conditions although generally lower variability.

Result Analysis

- Does LLM size matter?
  - This is analysed using the different sizes of Flan-T5 models used.
  - In general, larger models show higher effectiveness.
  - For pointwise there is an exception. Improvements are observed when passing from FlanT5-large to FlanT5-XL. However when using FlanT5-XXL both decreased effectiveness and large variance in effectiveness across the prompt variations are observed.

Result Analysis

- Does the LLM backbone matter?
  - Compares three backbones which are comparable size across the results from prompts
    - FlanT5-XXL (11.3B)
    - Mistral-7B
    - Llama3-8B
  - Llama3 and FlanT5 outperform Mistral based rankers.
  - Llama3 and FlanT5 have overall similar effectiveness, though on COVID Llama3-based rankers consistently outperform those with FlanT5.
  - Mistral-based rankers exhibit larger variance in effectiveness due to prompt variations than rankers based on the other backbones.

Limitations

- Query latency is not considered as the main focus is on prompt effectiveness as some long prompts are used.
- Even though 1248 is a large number, many more prompt variations could have been designed and investigated. Due to the limitation in GPU resources and computational costs, it was limited.
- This research is restricted to non commercial models due to cost limitations.

Conclusion

- Analysis revealed that ranking effectiveness varies considerably across different implementations of prompt components.
- Optimal prompt wording showed variability depending on the ranking method, dataset, and LLM backbone employed, suggesting that automatic prompt optimization, tailored to specific ranking methods and datasets, may be more effective than manual prompt engineering for optimizing ranking performance.

References

1) Ma, X., Zhang, X., Pradeep, R., Lin, J.: Zero-shot listwise document reranking with a large language model. arXiv preprint arXiv:2305.02156 (2023).
2) Sun, W., Yan, L., Ma, X., Ren, P., Yin, D., Ren, Z.: Is chatgpt good at search? investigating large language models as re-ranking agent. arXiv preprint arXiv:2304.09542 (2023).
3) Tang, R., Zhang, X., Ma, X., Lin, J., Ture, F.: Found in the middle: Permutation self-consistency improves listwise ranking in large language models. arXiv preprint arXiv:2310.07712 (2023).
4) Zhuang, H., Qin, Z., Hui, K., Wu, J., Yan, L., Wang, X., Bendersky, M.: Beyond yes and no: Improving zero-shot pointwise llm rankers via scoring fine-grained relevance labels. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2024).
5) Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Shen, J., Liu, T., Liu, J., Metzler, D., Wang, X., et al.: Large language models are effective text rankers with pairwise ranking prompting. arXiv preprint arXiv:2306.17563 (2023).
6) Ma, X., Zhang, X., Pradeep, R., Lin, J.: Zero-shot listwise document reranking with a large language model. arXiv preprint arXiv:2305.02156 (2023).
7) Zhuang, S., Zhuang, H., Koopman, B., Zuccon, G.: A setwise approach for effective and highly efficient zero-shot ranking with large language models. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 38–47 (2024).
8) Kim, J., Park, S., Jeong, K., Lee, S., Han, S.H., Lee, J., Kang, P.: Which is better? exploring prompting strategy for llm-based metrics. arXiv preprint arXiv:2311.03754 (2023)

References

9) Thomas, P., Spielman, S., Craswell, N., Mitra, B.: Large language models can accurately predict searcher preferences. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1930–1940 (2024).
10) Kamruzzaman, M., Kim, G.L.: Prompting techniques for reducing social bias in llms through system 1 and system 2 cognitive processes. arXiv preprint arXiv:2404.17218 (2024)
11) Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., Stanovsky, G.: State of what art? a call for multi-prompt llm evaluation. Transactions of the Association for Computational Linguistics 12, 933–949 (2024)
12) Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q.V., Zhou, D., Chen, X.: Large language models as optimizers. arXiv preprint arXiv:2309.03409 (2023).
13) Lin, J., Ma, X., Lin, S.C., Yang, J.H., Pradeep, R., Nogueira, R.: Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2356–2362 (2021).
14) Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. Journal of Machine Learning Research 25(70), 1–53 (2024)
15) Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
16) AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/ main/MODEL_CARD.md