UNIVERSIDADE DA CORUÑA

Limitations of Automatic Relevance Assessments with Large Language Models for Fair and Reliable Retrieval Evaluation

David Otero, Javier Parapar, Álvaro Barreiro

IRLab, CITIC, Universidade da Coruña, Spain

Motivation

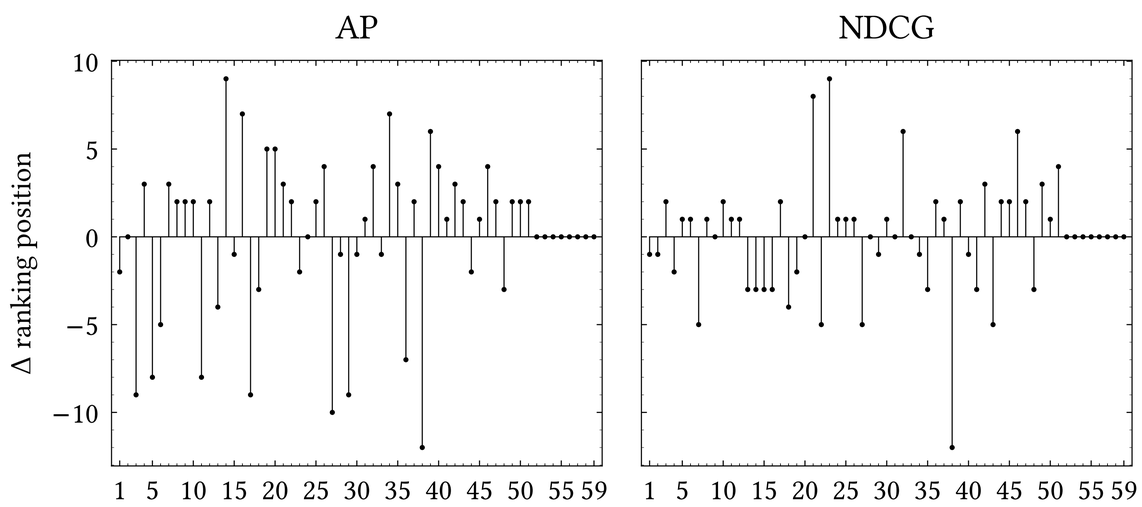
- Offline evaluation of IR systems relies on human assessments that are expensive and time-consuming to obtain [2].
- LLMs have shown promising results in making assessments for IR evaluation in terms of overall ranking correlation [3], but these correlations are not informative about top-performing systems and ignore if statistical significances are preserved.
- We investigate the limitations of LLM-generated judgements to evaluate IR systems in terms of correlation among top-performing systems and preservation of statistical significant differences.

Experimental Setup

- We use the data from the SynDL dataset, which includes judgements generated with GPT4 for the DL-2019, DL-2020, DL-2021, DL-2022, and DL-2023 tracks of TREC [1].
- We study the ranking of systems correlation among top ranks, using τ_{AP} [5] and RBO [4], which are top-weighted correlations that penalize more misses at the top.
- We look at how well the LLM-generated relevance assessments preserve statistical significance detected under human judgements.

Correlation among Top Ranks

- **Top-weighted correlations have lower figures than overall correllation** (more details on the paper), showing that LLM-generated judgement might not be still able to distinguish the best systems.
- We show on the right, for the particular case of DL-2020, the change in terms of ranking position that each run suffers when evaluated with LLM-generated relevance assessments compared to the original human-generated ones.
- We observe that only the tail of the original ranking is preserved, while the **top systems are shuffled**.



Preservation of Statistical Significance

Metric		Dataset				
		DL-2019	DL-2020	DL-2021	DL-2022	DL-2023
AP	TP	68%	88%	88%	89%	91%
	FN	32%	12%	12%	11%	9%
	TN	57%	62%	50%	51%	28%
	FP	43%	38%	50%	49%	72%
NDCG	TP	83%	94%	91%	92%	93%
	FN	17%	6%	9%	8%	7%
	TN	58%	63%	64%	55%	33%
	FP	42%	37%	36%	45%	67%

- We assume that significance decision derived from human judgements are the ground truth, and we evaluate how LLM-generated judgements are able to preseve these differences.
- We computed the true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP).
- We employed a two-sided Wilcoxon Signed Rank test to determine statistical significance.
- We observe that the LLM-generated judgements incur in a **high rate of false positives**, which means that they tend to indicate significance when there is none.

Conclusions

- Making relevance judgements is the most resource-intensive task when creating new IR test collections, and LLMs offer opportunities to alleviate this cost.
- In terms of distinguishing the best systems, we showed that correlations degrade as we focus more and more on the best systems.
- In terms of statistical significance, we found that LLM-generated judgements incur in a exeedingly high rate of false positives.

References

[1] H. A. Rahmani, X. Wang, E. Yilmaz, N. Craswell, B. Mitra, and P. Thomas. Syndl: A large-scale synthetic test collection for passage retrieval, 2024.

[2] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010.

[3] S. Upadhyay, R. Pradeep, N. Thakur, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, and J. Lin. A large-scale study of relevance assessments with large language models: An initial look, 2024.

[4] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4), 2010.

[5] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 587–594, New York, NY, USA, 2008. Association for Computing Machinery.

IRLAB INFORMATION RETRIEVAL LAB · citic · UNIVERSIDADE DA CORUÑA · MINISTERIO DE CIENCIA E INNOVACIÓN · Financiado por la Unión Europea NextGenerationEU · Plan de Recuperación, Transformación y Resiliencia · AGENCIA ESTATAL DE INVESTIGACIÓN · XUNTA DE GALICIA