UNIVERSITY OF AMSTERDAM

Opportunities and Challenges of LLMs in Information Retrieval

Chuan Meng

University of Amsterdam

21st October 2024

About me

As of Oct 2024, I have authored 15 papers 230 citations (Google Scholar) with an H-index of 7

Chuan Meng

- Final-year PhD student at the University of Amsterdam
  - Supervisor: Maarten de Rijke, Mohammad Aliannejadi
- Applied Scientist Intern at Amazon (London)
  - Manager: Gabriella Kazai, mentor: Francesco Tonolini

- Research directions:
  - Conversational agents
    - (Proactive) conversational search,
    - Knowledge-grounded dialogue systems
  - Neural ranking
    - Generative retrieval,
    - LLM-based re-ranking,
  - Automatic evaluation
    - Query performance prediction (QPP),
    - LLM-based relevance judgment prediction

Background

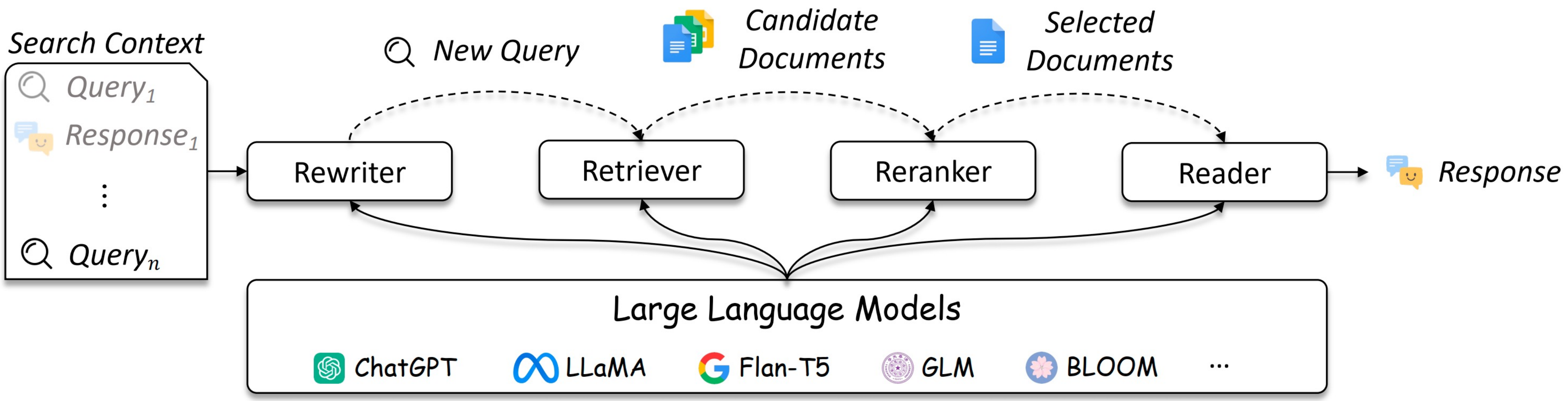
- Large language models (LLMs) have remarkable language understanding, generation, generalization, and reasoning abilities

Which is the best university in the UK for Information Retrieval research?

The **University of Glasgow** is often regarded as the leading institution for Information Retrieval (IR) research in the UK. Its **Glasgow Information Retrieval Group** is internationally recognised for its pioneering work in areas such as **neural IR models, query performance prediction, large-scale retrieval systems**, and **evaluation methodologies**. The group has made significant contributions to the field, including developing the widely-used **Terrier** IR platform.

Background

- LLMs in information retrieval (IR)
  - Opportunities
    - LLMs directly as task solvers (e.g., retrievers/re-rankers)
    - LLMs for data augmentation (e.g., training retrievers/re-rankers)
    - LLMs for automatic evaluation (e.g., generating relevance judgments)
  - Challenges
    - Low efficiency
    - Unfaithful generation
    - …



Background

- LLMs in information retrieval (IR)
  - Opportunities
    - LLMs directly as task solvers (e.g., retrievers/re-rankers) [1,2,3]
    - LLMs for data augmentation (e.g., training retrievers/re-rankers) [4,5]
    - LLMs for automatic evaluation (e.g., generating relevance judgments) [6,7]
  - Challenges
    - Low efficiency [8]
    - Unfaithful generation
    - …

[1] Generative Retrieval with Few-shot Indexing. arXiv 2024.
[2] LLM-based Retrieval and Generation Pipelines for TREC Interactive Knowledge Assistance Track (iKAT) 2023. TREC 2023.
[3] System Initiative Prediction for Multi-turn Conversational Information Seeking. CIKM 2023.
[4] Expand, Highlight, Generate: RL-driven Document Generation for Passage Reranking. EMNLP 2023.
[5] Self-seeding and Multi-intent Self-instructing LLMs for Generating Intent-aware Information-Seeking dialogs. arXiv 2024.
[6] Query Performance Prediction using Relevance Judgments Generated by Large Language Models. arXiv 2024.
[7] Can We Use Large Language Models to Fill Relevance Judgment Holes? LLM4Eval 2024.
[8] Ranked List Truncation for Large Language Model-based Re-Ranking. SIGIR 2024.

Background

- LLMs in information retrieval (IR)
  - Opportunities
    - LLMs directly as task solvers (e.g., retrievers/re-rankers) [**1**,2,3]
    - LLMs for data augmentation (e.g., training retrievers/re-rankers) [4,5]
    - LLMs for automatic evaluation (e.g., generating relevance judgments) [**6**,7]
  - Challenges
    - Low efficiency [**8**]
    - Unfaithful generation
    - …

[1] Generative Retrieval with Few-shot Indexing. arXiv 2024.
[2] LLM-based Retrieval and Generation Pipelines for TREC Interactive Knowledge Assistance Track (iKAT) 2023. TREC 2023.
[3] System Initiative Prediction for Multi-turn Conversational Information Seeking. CIKM 2023.
[4] Expand, Highlight, Generate: RL-driven Document Generation for Passage Reranking. EMNLP 2023.
[5] Self-seeding and Multi-intent Self-instructing LLMs for Generating Intent-aware Information-Seeking dialogs. arXiv 2024.
[6] Query Performance Prediction using Relevance Judgments Generated by Large Language Models. arXiv 2024.
[7] Can We Use Large Language Models to Fill Relevance Judgment Holes? LLM4Eval 2024.
[8] Ranked List Truncation for Large Language Model-based Re-Ranking. SIGIR 2024.

Outline

- ❑ Study 1: using LLMs as few-shot generative retriever [10 min]
- ❑ Study 2: using LLMs as relevance judgment and query performance predictor [10 min]
- ❑ Study 3: improve the efficiency of LLM-based re-rankers [15 min]
- ❑ Conclusion [5 min]

Outline

- ❑ **Study 1: using LLMs as few-shot generative retriever** [10 min]
- ❑ Study 2: using LLMs as relevance judgment and query performance predictor [10 min]
- ❑ Study 3: improve the efficiency of LLM-based re-rankers [15 min]
- ❑ Conclusion [5 min]

Generative Retrieval with Few-shot Indexing

Arian Askari*, **Chuan Meng***, Mohammad Aliannejadi, Zhaochun Ren,
Evangelos Kanoulas, Suzan Verberne

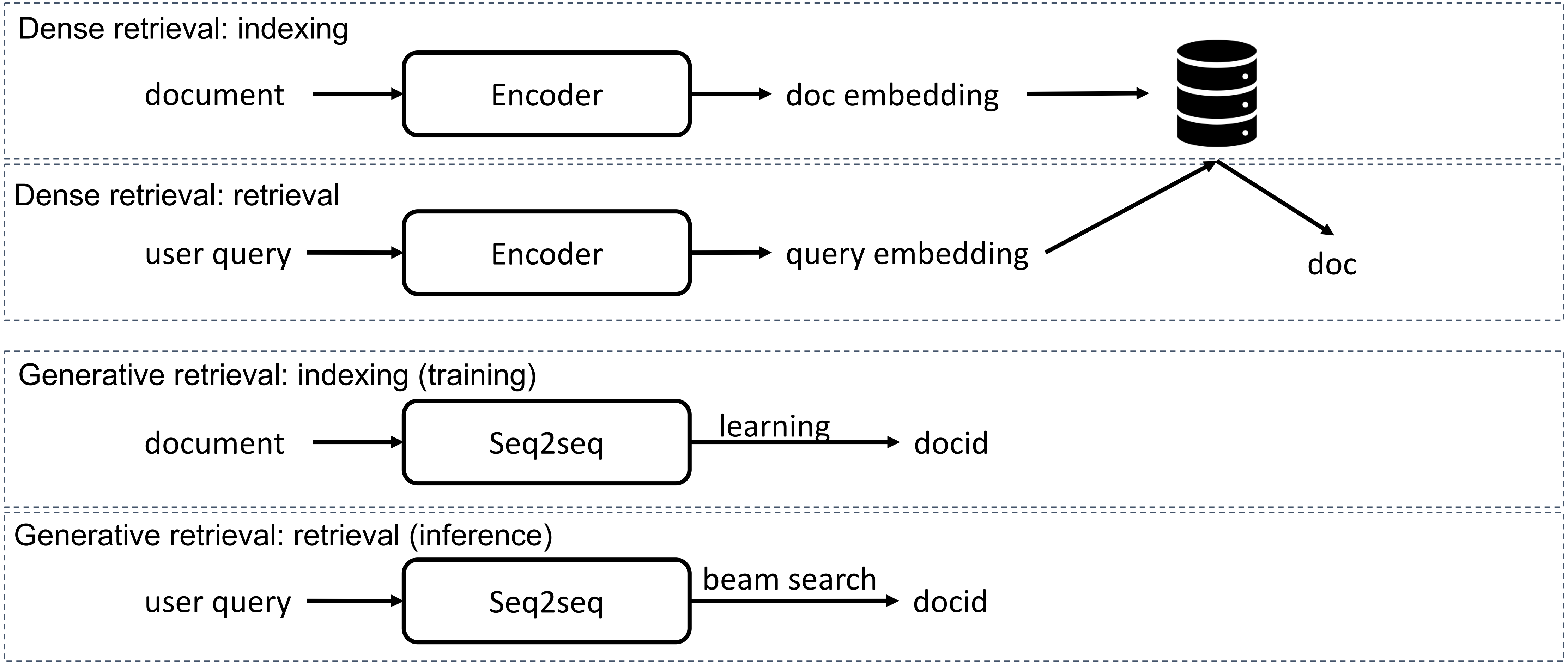
University of Amsterdam
Universiteit Leiden
The Netherlands

arXiv 2024

* denotes co-first authors

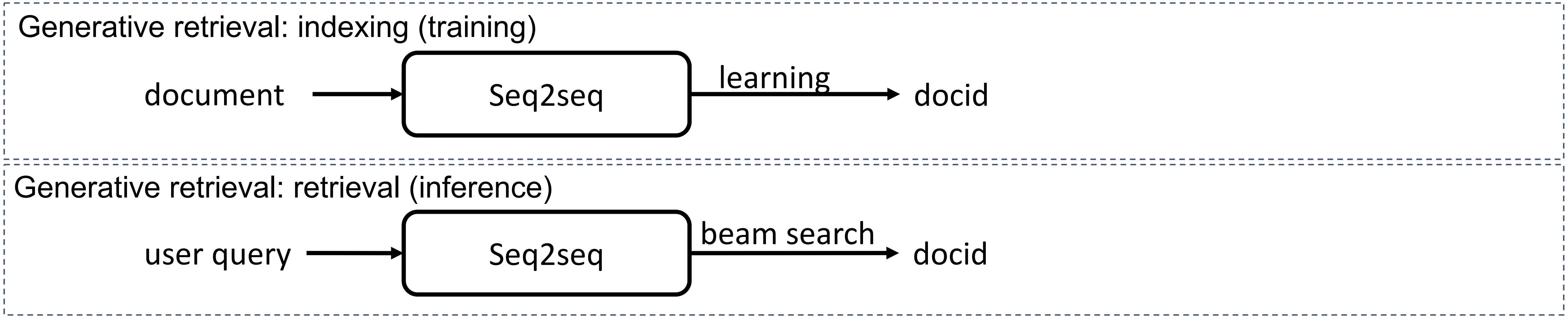
Background—generative retrieval

- Generative retrieval consolidates indexing and retrieval into a single model
  - Indexing (training) trains a seq2seq model to map document text to its docid
  - Retrieval (inference) feeds the model a query text to generate relevant docids



Motivation

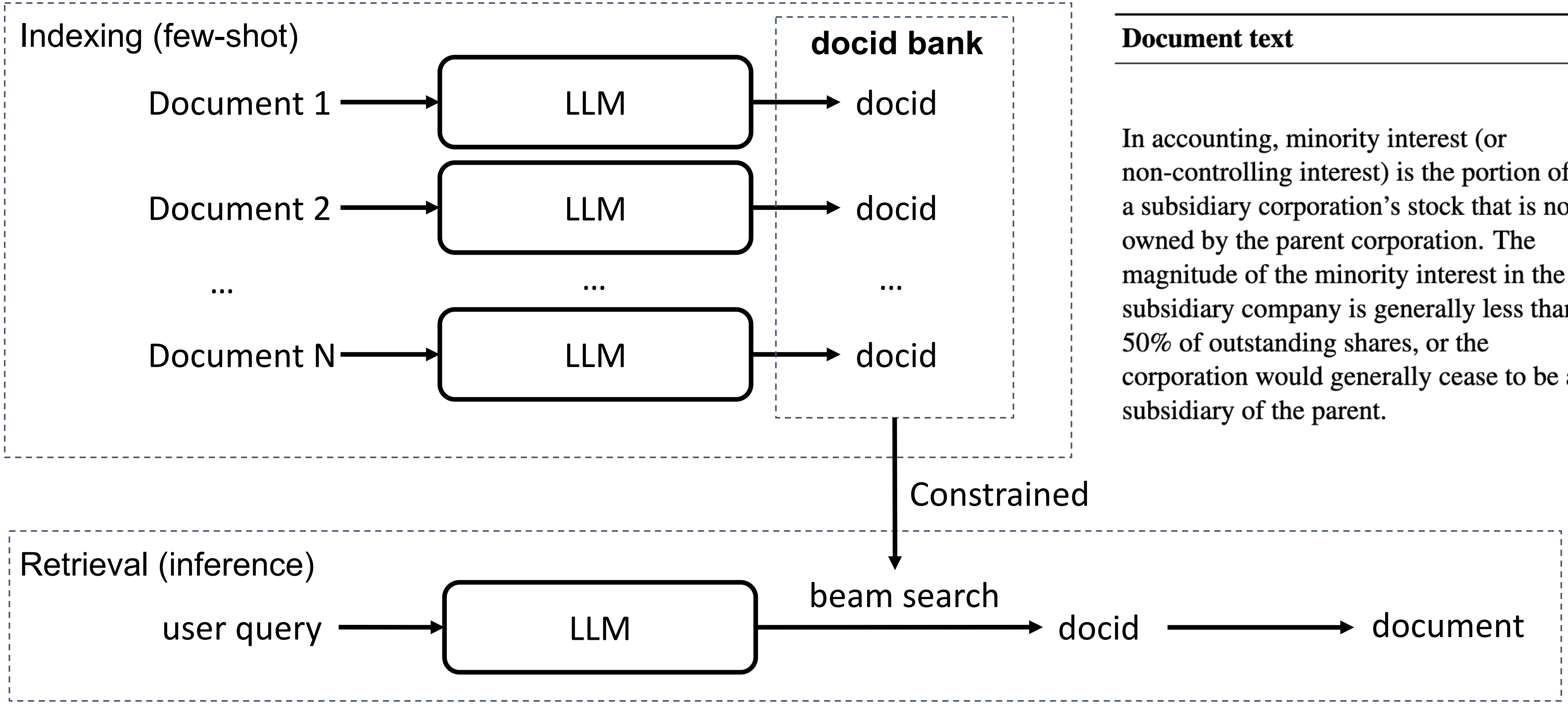
- Previous studies typically rely on **training-based indexing**
  - high training overhead
    - the authors of GenRET indicated the training took 7 days on 100 A100 GPUs [1]
  - under-utilization of the pre-trained knowledge of LLMs
  - hard to adapt to a dynamic document corpus



[1] Sun et al. Learning to Tokenize for Generative Retrieval. NeurIPS 2023.

Methodology

- We propose a **few-shot** indexing-based **g**enerative **r**etrieval framework (Few-shot GR)



Document text	docids
In accounting, minority interest (or non-controlling interest) is the portion of a subsidiary corporation's stock that is not owned by the parent corporation. The magnitude of the minority interest in the subsidiary company is generally less than 50% of outstanding shares, or the corporation would generally cease to be a subsidiary of the parent.	minority-interest-accounting
	non-controlling-interest-accounting
	minority-interest-definition
	minority-interest-calculation
	minority-interest-significance

Experiments

- Experiments on NQ320K show Few-shot GR
  - achieves superior performance to SOTA baselines that require heavy training
  - is much more efficient than SOTA baselines

Method	Recall@1	Recall@10	MRR@100
BM25	29.7	60.3	40.2
DocT5Query	38.0	69.3	48.9
DPR	50.2	77.7	59.9
ANCE	50.2	78.5	60.2
SentenceT5	53.6	83.0	64.1
GTR-base	56.0	84.4	66.2
SEAL	59.9	81.2	67.7
DSI	55.2	67.4	59.6
NCI	66.4	85.7	73.6
DSI-QG	63.1	80.7	69.5
DSI-QG (InPars)	63.9	82.0	71.4
GenRET	68.1	**88.8**	<u>75.9</u>
TOME	66.6	–	–
GLEN	<u>69.1</u>	86.0	75.4
Few-Shot GR	**70.1**	<u>87.6</u>	**77.4**

Method	Indexing (hr)	Retrieval (ms)
DSI-QG	240	72
GenRET	≈16,800	72
Few-Shot GR	37	98

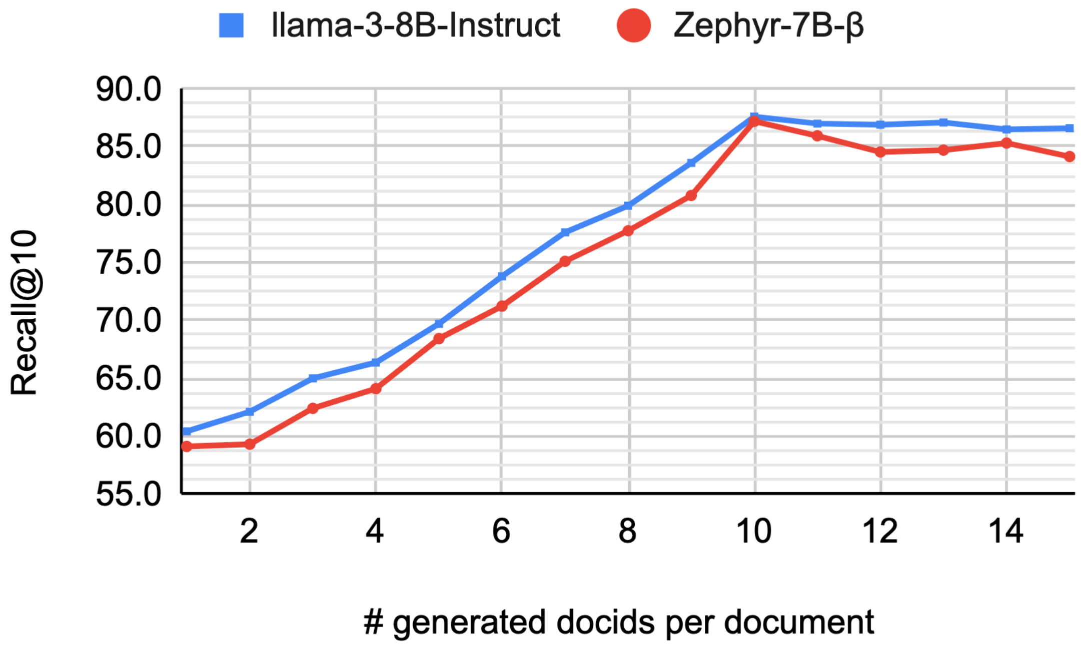
The authors of GenRET indicated it took 7 days on 100 A100 GPUs ≈16,800 hours on a single A100 GPU

Experiments

- Selecting a generally stronger LLM leads to better performance

Method	Recall@1	Recall@10	MRR@100
T5-base	52.4	66.4	55.8
Zephyr-7B-β	69.9	87.2	**77.8**
llama-3-8B-Instruct	**70.1**	**87.6**	77.4

- Performance improves as generating more docids per document during indexing



Conclusions and Future Work

- Contributions
  - Propose Few-shot GR, a new generative retrieval paradigm
    - performing indexing only by prompting an LLM
    - achieving superior performance to SOTA baselines that require heavy training
    - significantly reducing indexing overhead

- Future work
  - Test Few-shot GR on a document corpus with millions of documents

Q & A

Outline

- ❑ Study 1: using LLMs as few-shot generative retriever [10 min]
- ❑ **Study 2: using LLMs as relevance judgment and query performance predictor** [10 min]
- ❑ Study 3: improve the efficiency of LLM-based re-rankers [15 min]
- ❑ Conclusion [5 min]

University of Amsterdam University of Waterloo Universiteit Leiden The Netherlands

Query Performance Prediction using Relevance Judgments Generated by Large Language Models

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, Maarten de Rijke

arXiv 2024

Motivation

- Prompting open-source LLMs results in limited performance in predicting relevance judgments

LLM	TREC-DL 19	TREC-DL 20	TREC-DL 21	TREC-DL 22
	κ	κ	κ	κ
GPT-3.5 (text-davinci-003) [32]	-	-	0.260	-
LLaMA-7B (few-shot)	-0.001	-0.003	0.003	-0.010
Llama-3-8B (few-shot)	0.018	0.027	0.021	-0.035
Llama-3-8B-Instruct (few-shot)	0.315	0.227	0.238	0.049

Methodology

- Fine-tuning open-source LLMs for generating relevance judgments
  - LLMs: LLaMA-7B, Llama-3-8B, and Llama-3-8B-Instruct
  - Fine-tuning method: QLoRA, a parameter-efficient fine-tuning method
  - Training data: human-labeled relevance judgments of MS MARCO

Instruction: Please assess the relevance of the provided passage to the following question. Please output "Relevant" or "Irrelevant".
Question: {question}
Passage: {passage}
Output: Relevant/Irrelevant

Experiments

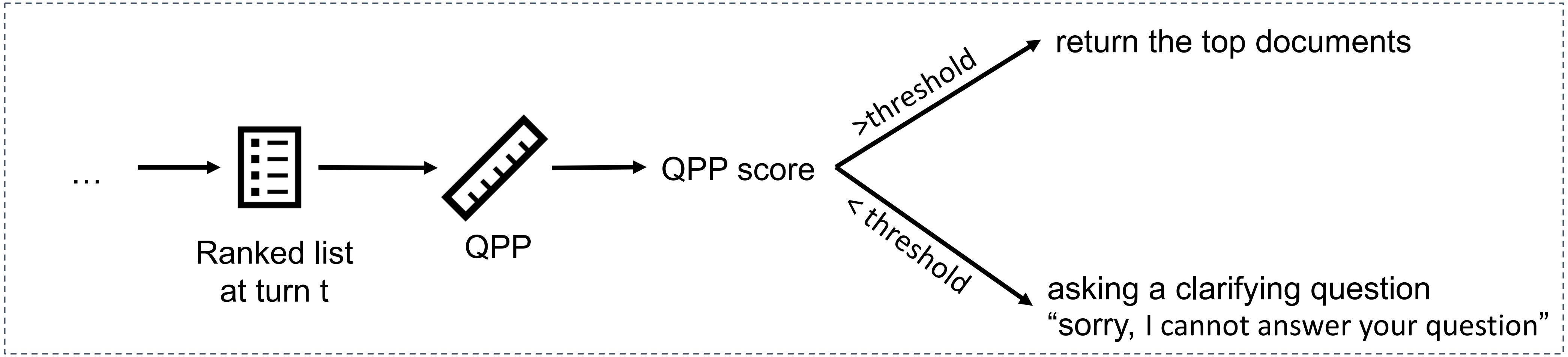
- Fine-tuned LLMs outperform
  - their counterparts using few-shot prompting
  - GPT-3.5

LLM	TREC-DL 19 κ	TREC-DL 20 κ	TREC-DL 21 κ	TREC-DL 22 κ
GPT-3.5 (text-davinci-003) [32]	-	-	0.260	-
LLaMA-7B (few-shot)	-0.001	-0.003	0.003	-0.010
Llama-3-8B (few-shot)	0.018	0.027	0.021	-0.035
Llama-3-8B-Instruct (few-shot)	0.315	0.227	0.238	0.049
LLaMA-7B (fine-tuned)	0.258	0.238	0.333	0.038
Llama-3-8B (fine-tuned)	0.381	**0.342**	0.347	**0.082**
Llama-3-8B-Instruct (fine-tuned)	**0.397**	0.316	**0.418**	0.066

Background—query performance prediction

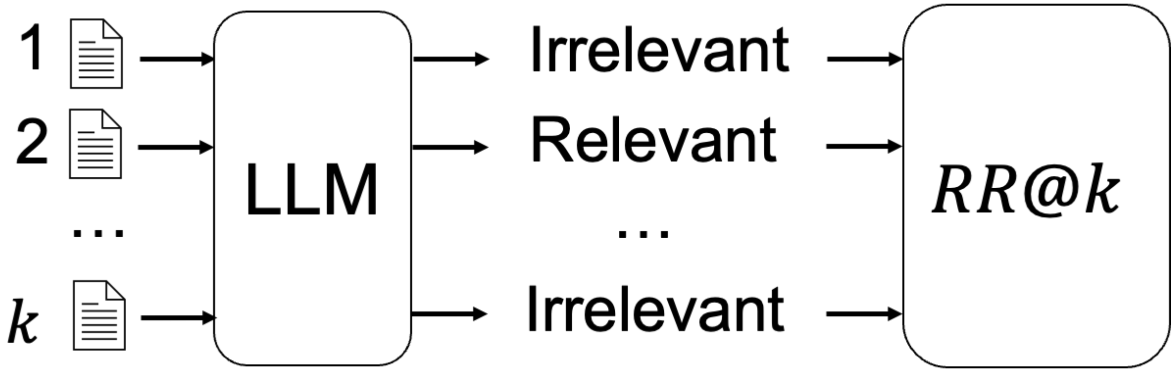
- Query performance prediction (QPP)
    - Predicts retrieval quality of search system for query without human-labeled relevance judgments

- QPP benefits a variety of applications, e.g., action prediction in conversational search

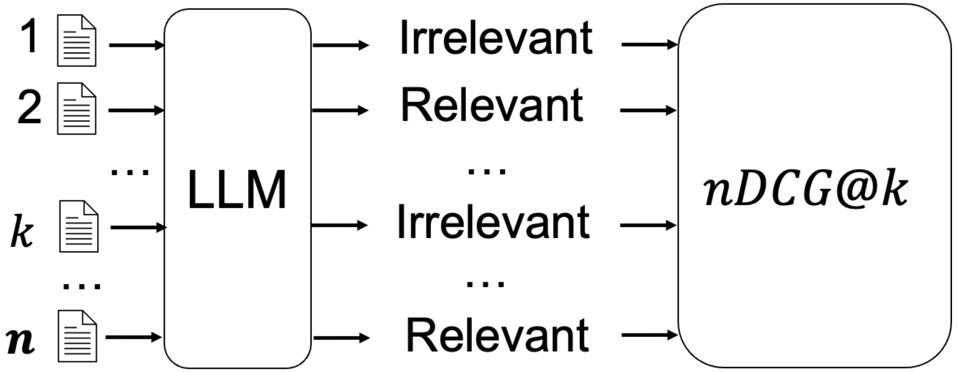


Methodology

- Propose QPP-GenRE, which predicts IR measures using LLM-generated judgments
  - devise an approximation strategy for predicting a metric considering recall
    - only judges the top $\boldsymbol{n}$ items in a ranked list, where $\boldsymbol{n} \ll$ # documents in the corpus



Predicting a precision-based metric



Predicting a metric considering recall

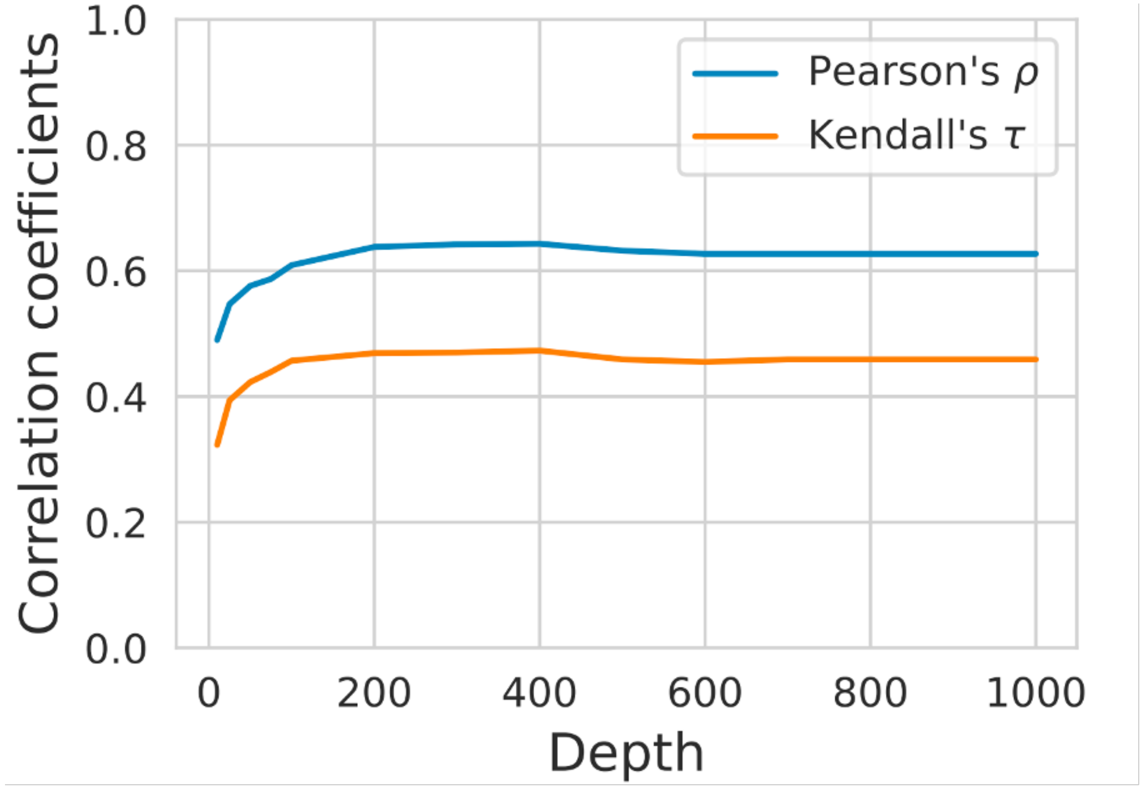
Experiments

- QPP-GenRE with fine-tuned LLMs achieves SOTA QPP quality

QPP method	TREC-DL 19		TREC-DL 20	
	P-ρ	K-τ	P-ρ	K-τ
Clarity	0.091	0.056	0.358*	0.250*
WIG	0.520*	0.331*	0.615*	0.423*
NQC	0.468*	0.300*	0.508*	0.401*
σ_{max}	0.478*	0.327*	0.529*	0.440*
n($\sigma_{x\%}$)	0.532*	0.311*	0.622*	0.443*
SMV	0.376*	0.271*	0.463*	0.383*
UEF(NQC)	0.499*	0.322*	0.517*	0.356*
RLS(NQC)	0.469*	0.169	0.522*	0.376*
QPP-PRP	0.321	0.181	0.189	0.157
NQA-QPP	0.210	0.147	0.244	0.210*
BERTQPP	0.458*	0.207	0.426*	0.300*
qppBERT-PL	0.171	0.175	0.410*	0.279*
M-QPPF	0.404*	0.254*	0.435*	0.297*
QPP-LLM (few-shot)	-0.024	-0.031	0.167	0.138
QPP-LLM (fine-tuned)	0.313*	0.215	0.309*	0.254*
QPP-GenRE (n = 200)	**0.724**†*	0.474†*	**0.638**†*	**0.469**†*
QPP-GenRE (n = 10)	0.605*	**0.482***	0.490*	0.323*
QPP-GenRE (n = 100)	0.712*	0.472*	0.609*	0.457*
QPP-GenRE (n = 1,000)	0.715*	0.477*	0.627*	0.459*

Experiments

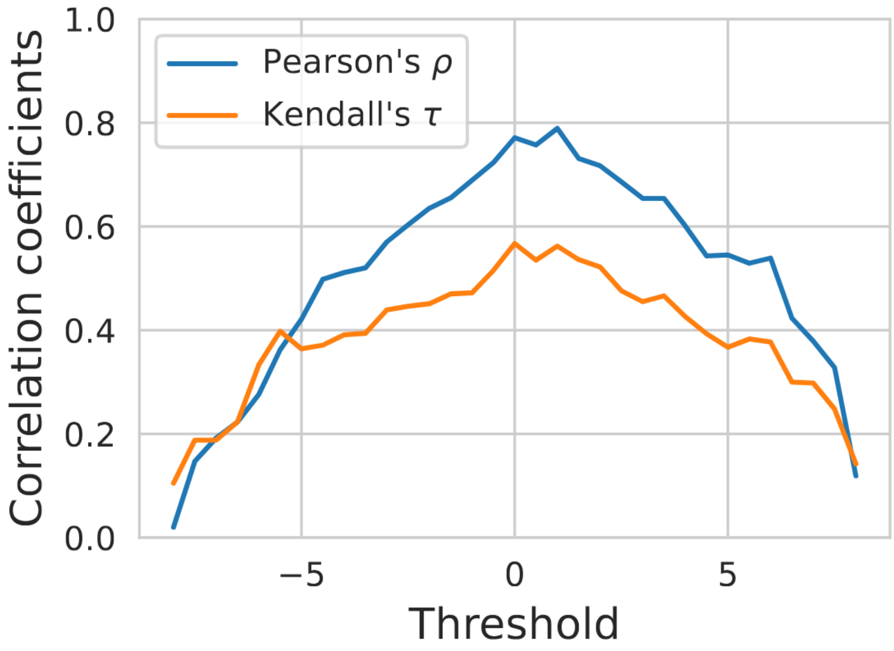
- Judging up to 100–200 items in a ranked list suffices for predicting nDCG@10



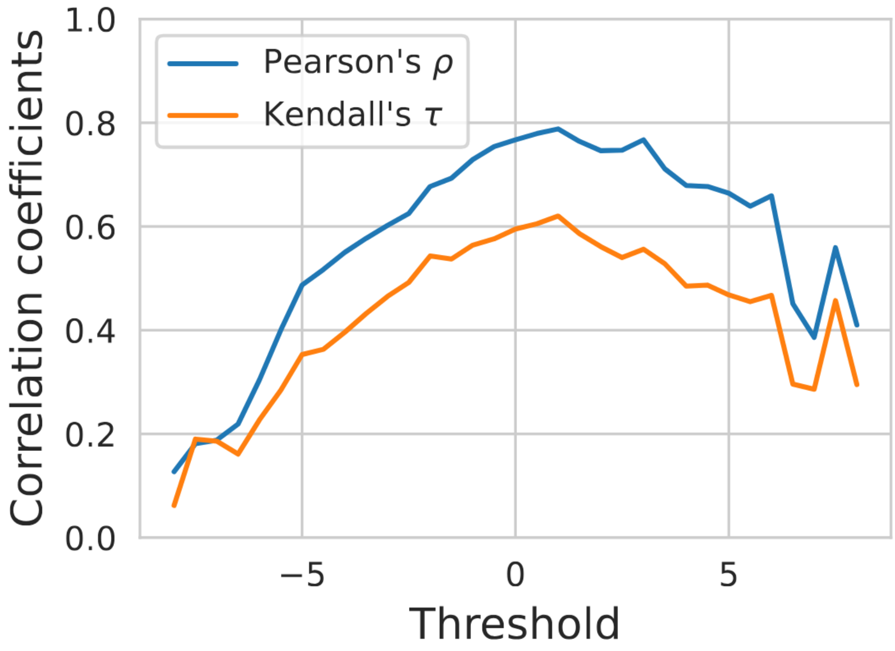
QPP quality of predicting BM25's nDCG@10 w.r.t. judging depth (n)

Experiments

- Integrating QPP-GenRE with RankLLaMA, an LLM-based point-wise re-ranker
  - Setting a threshold to convert a re-ranking score into a judgment label
  - A tuned threshold results in high QPP quality



(a) BM25 on TREC-DL 19



(b) BM25 on TREC-DL 20

Conclusion

- Contributions
  - Fine-tune open-source LLMs for generating relevance judgments
  - Propose a new QPP framework, QPP-GenRE, which predicts IR metrics based on LLM-generated relevance judgments
    - Devise an approximation strategy for predicting a metric considering recall
  - QPP-GenRE achieves state-of-the-art QPP quality
  - The data, code and fine-tuned checkpoints of LLMs are open-sourced
    https://github.com/ChuanMeng/QPP-GenRE

Q & A

QR code for the repo

Outline

- ❑ Study 1: using LLMs as few-shot generative retriever [10 min]
- ❑ Study 2: using LLMs as relevance judgment and query performance predictor [10 min]
- ❑ **Study 3: improve the efficiency of LLM-based re-rankers** [15 min]
- ❑ Conclusion [5 min]

UNIVERSITY OF AMSTERDAM

UNIVERSITY OF WATERLOO

Universiteit Leiden
The Netherlands

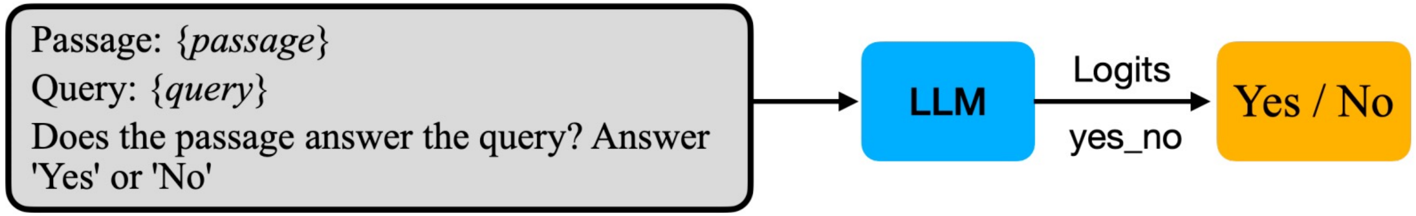
Ranked List Truncation for Large Language Model-based Re-Ranking

Chuan Meng, Negar Arabzadeh, Arian Askari, Mohammad Aliannejadi, Maarten de Rijke

The 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)

Background

- Large language models (LLMs) as text re-rankers
  - achieve state-of-the-art performance
  - hard to be applied in practice due to significant computational overhead
    - the average query latency (re-ranking 100 items per query) for Flan-t5-xxl (11B) is around 4 seconds, on a NVIDIA RTX A6000 GPU [1]



LLM-based re-ranker

[1] Zhuang et al. A Setwise Approach for Effective and Highly Efficient Zero-shot Ranking with Large Language Models. SIGIR 2024.

Motivation

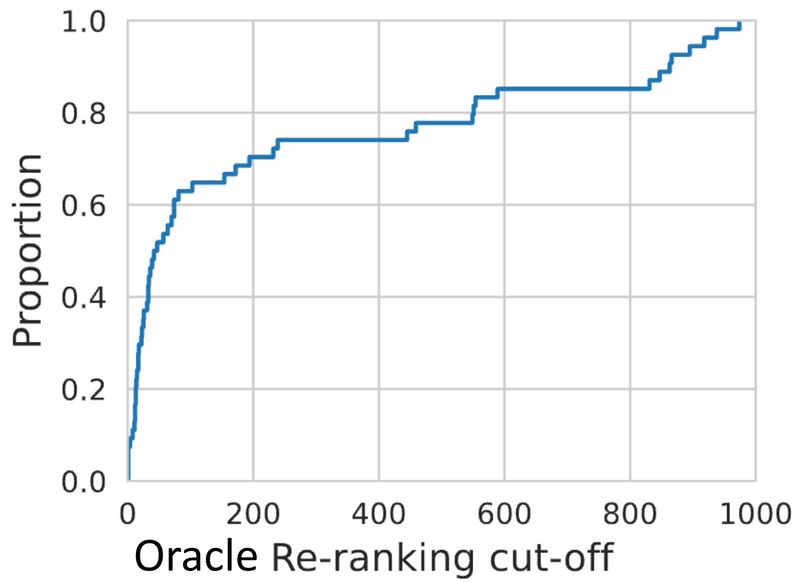
- Common practice: applying a fixed re-ranking cut-off to all queries (e.g., 100, 200, 1000)

- However,
  - a fixed re-ranking cut-off might lead to a waste of computational resources
  - individual queries might need a shorter or a longer list of re-ranking candidates

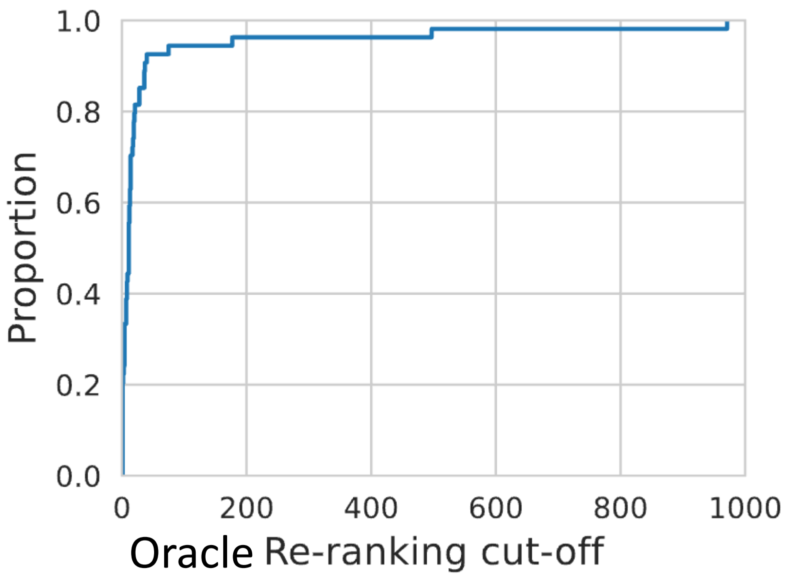
- We explore query-specific re-ranking cut-offs in the context of LLM-based re-ranking
  - Fixed cut-offs vs. query-specific cut-offs
  - How to predict query-specific cut-offs

Motivation (fixed cut-offs vs. query-specific cut-offs)

- Query-specific re-ranking cut-offs improve *efficiency*
  - Individual queries have different oracle cut-offs with a wide range
  - A deep fixed cut-off wastes computational resources
  - A shallow fixed cut-off hurts re-ranking quality for queries needing a deeper cut-off



Cumulative distribution function of oracle cut-offs for
BM25–RankLLaMA
TREC-DL 20

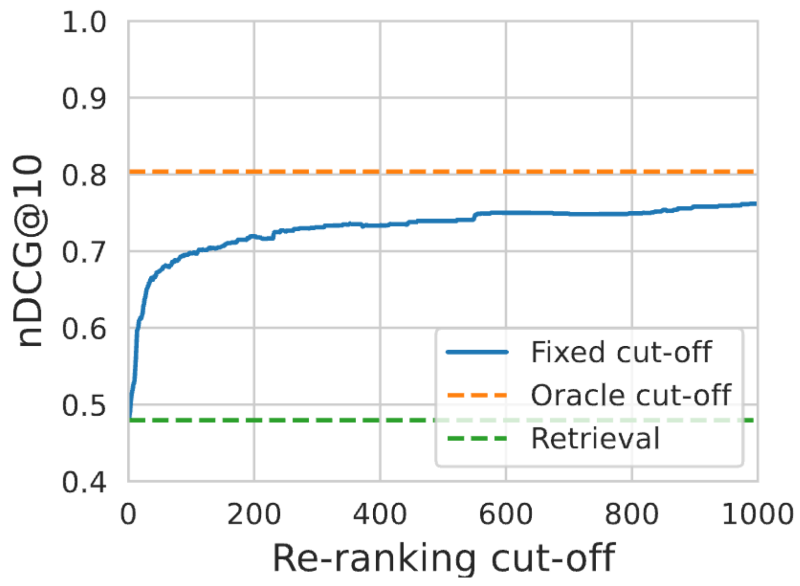


Cumulative distribution function of oracle cut-offs for
RepLLaMA–RankLLaMA
TREC-DL 20

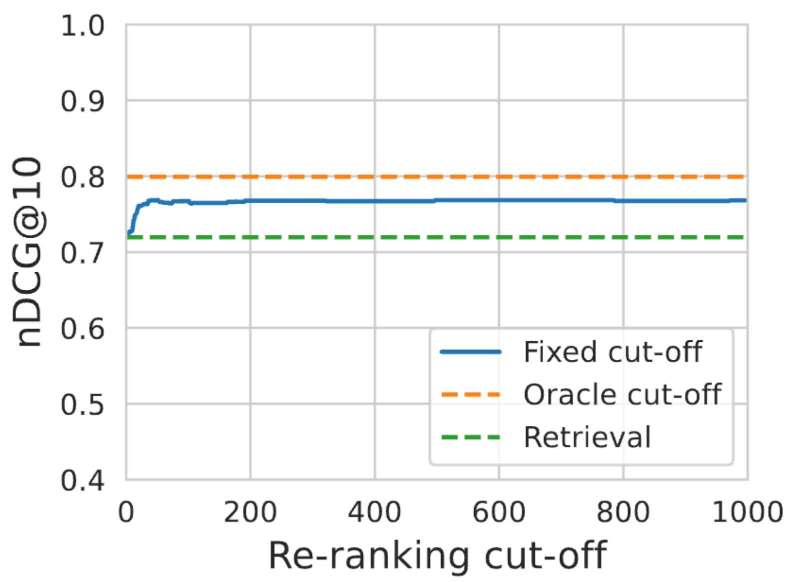
For a query, an oracle cut-off is the minimum re-ranking cutoff producing the highest nDCG@10 value

Motivation (fixed cut-offs vs. query-specific cut-offs)

- Query-specific re-ranking cut-offs improve *effectiveness*
  - Oracle cut-offs show statistically significant improvements over all fixed cut-offs
  - A deeper fixed cut-off
    - does not always result in improvement (consistent with [1])
    - even is detrimental to re-ranking quality (consistent with [1])



BM25–RankLLaMA
TREC-DL 20



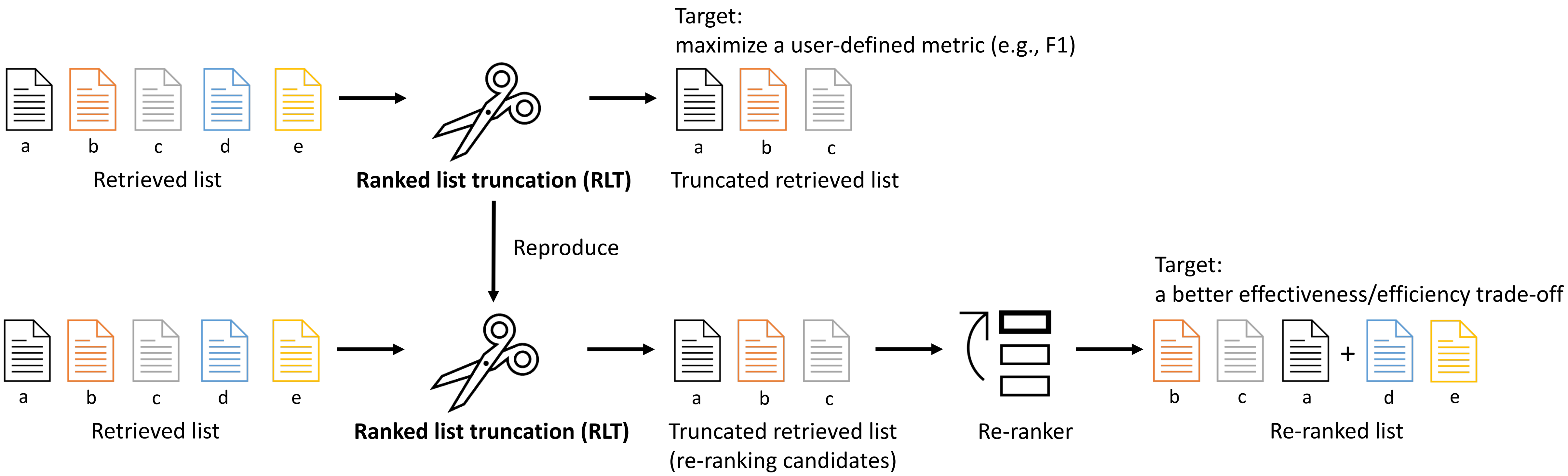
RepLLaMA–RankLLaMA
TREC-DL 20

[1] Zamani et al. Stochastic Retrieval-Conditioned Reranking. In ICTIR 2022.

Motivation (How to predict query-specific cut-offs)

- Ranked list truncation (RLT)
  - predicts how many items in a ranked list should be returned
  - optimizes the truncated ranked list regarding a user-defined metric (e.g., F1)
  - aids applications where reviewing returned items is costly, e.g., patent or legal search

- **We reproduce exiting RLT methods in the context of LLM-based re-ranking**



Reproducibility methodology

- *Do RLT methods generalize to the context of*
    - *(RQ1) LLM-based re-ranking with a lexical first-stage retriever?*
    - *(RQ2) LLM-based re-ranking with learned sparse or dense first-stage retrievers?*
    - *(RQ3) pre-trained language model-based re-ranking?*

Reproducibility methodology

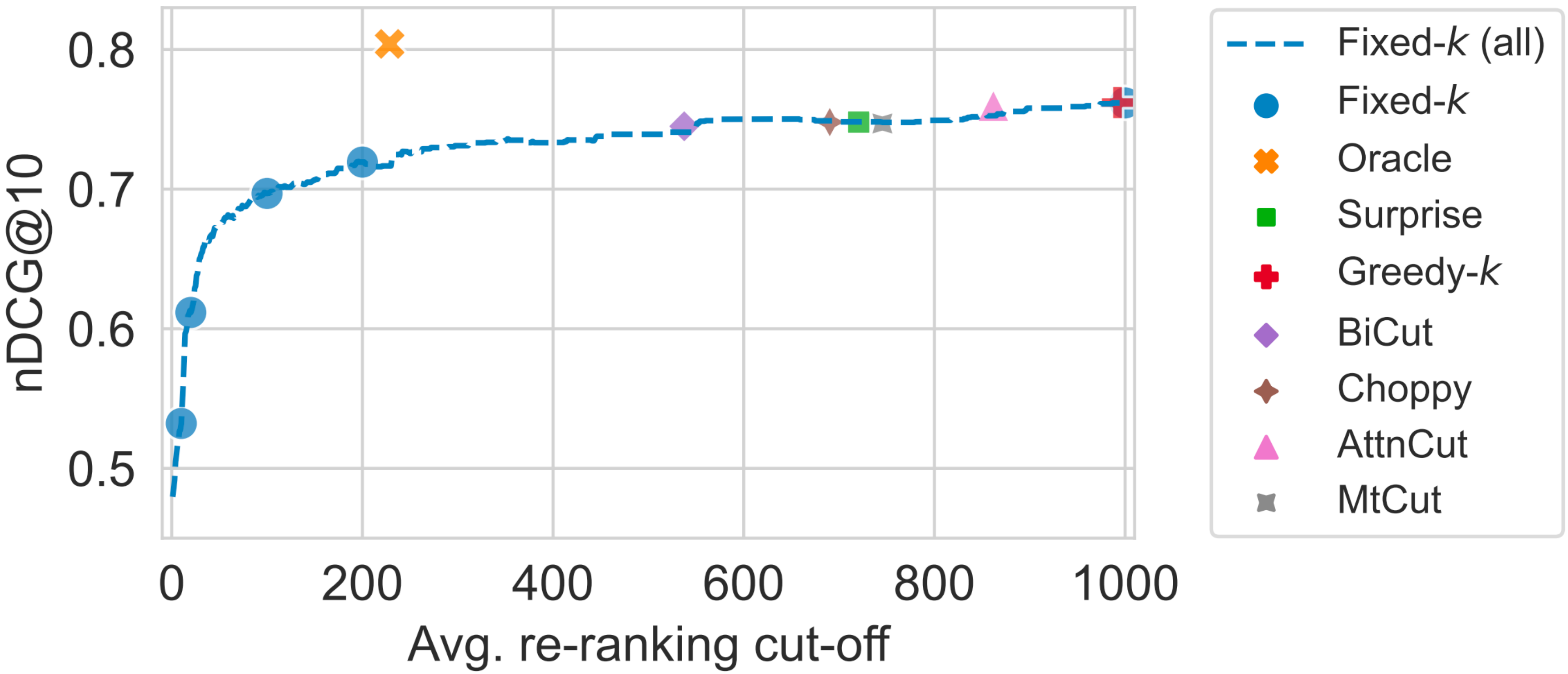
- Experimental settings:
  - 8 RLT methods

Method	Attribute 1
Fixed-k (10, 20, 100, 200, 1000)	Unsupervised
Greedy-k	Unsupervised
Surprise	Unsupervised
BiCut	Supervised
Choppy	Supervised
AttnCut	Supervised
MtCut	Supervised
LeCut	Supervised

  - Datasets:
    - TREC-DL 19, TREC-DL 20

Experiments

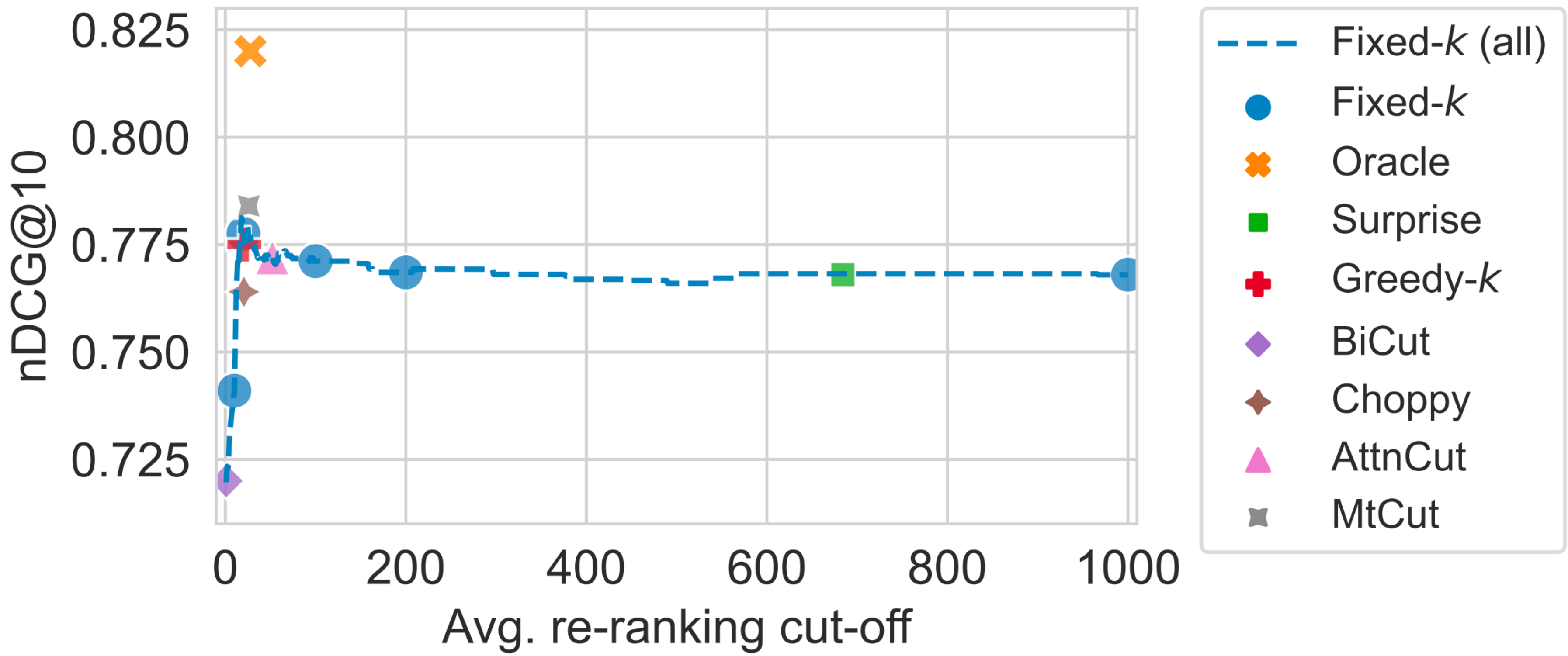
- RQ1: Do RLT methods generalize to the context of LLM-based re-ranking with a lexical first-stage retriever?
  - Fixed re-ranking depths can closely approximate supervised RLT methods' results
  - Supervised RLT methods do not show a clear advantage over fixed re-ranking depths



BM25–RankLLaMA
TREC-DL 20

Experiments

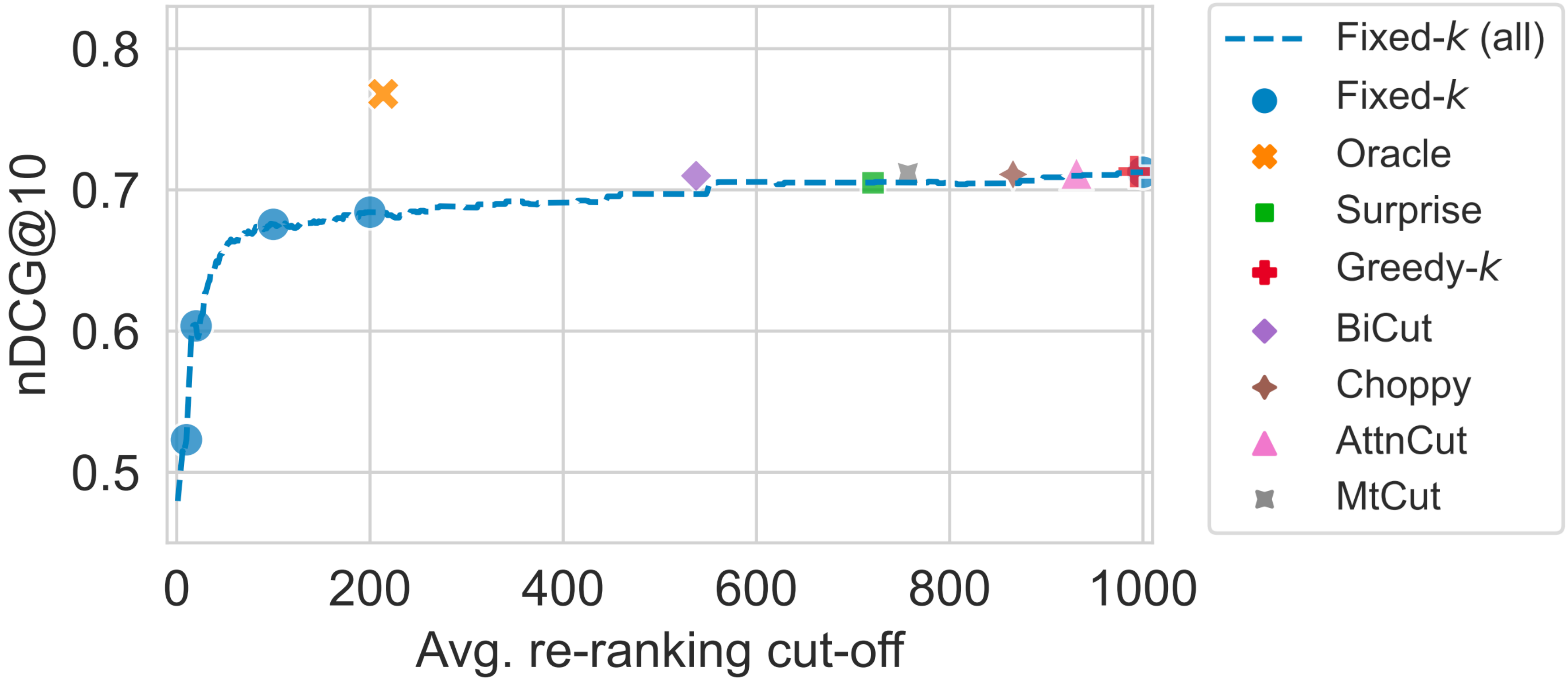
- RQ2: Do RLT methods generalize to the context of LLM-based re-ranking with learned sparse or dense first-stage retriever?
  - Supervised methods do not lead to significant improvement in terms nDCG@10
  - A fixed re-ranking depth of 20 achieves the best effectiveness/efficiency trade-off



Splade++–RankLLaMA
TREC-DL 20

Experiments

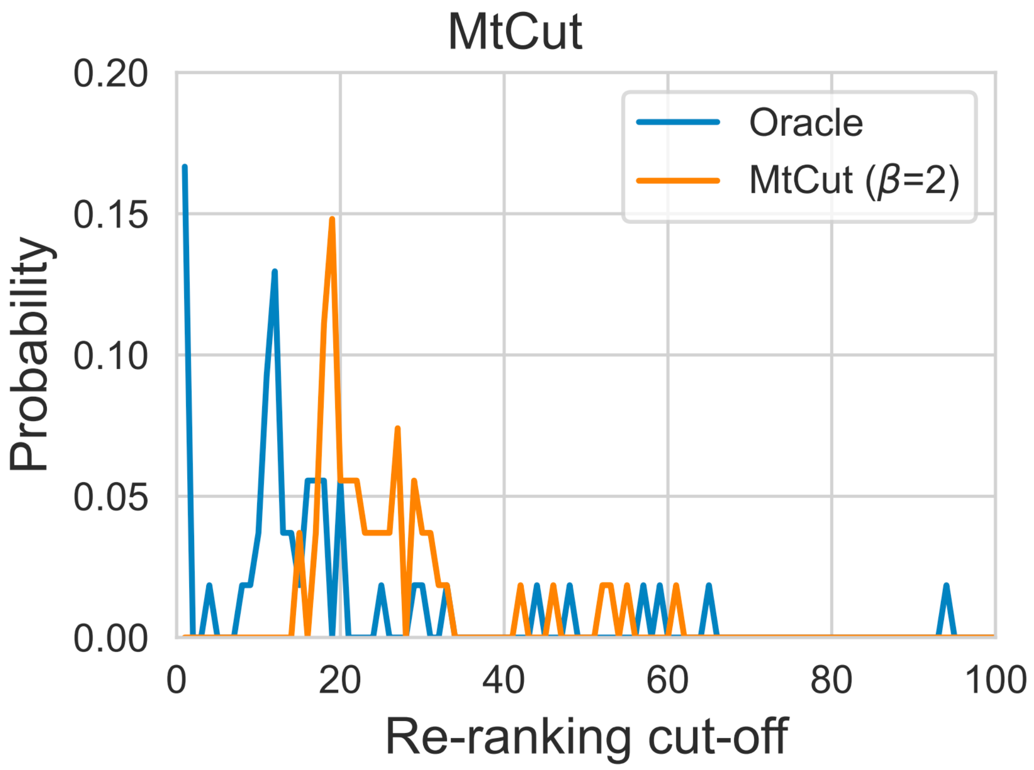
- RQ3: Do RLT methods generalize to the context of pre-trained language model-based re-ranking?
  - Results are similar to RQ1



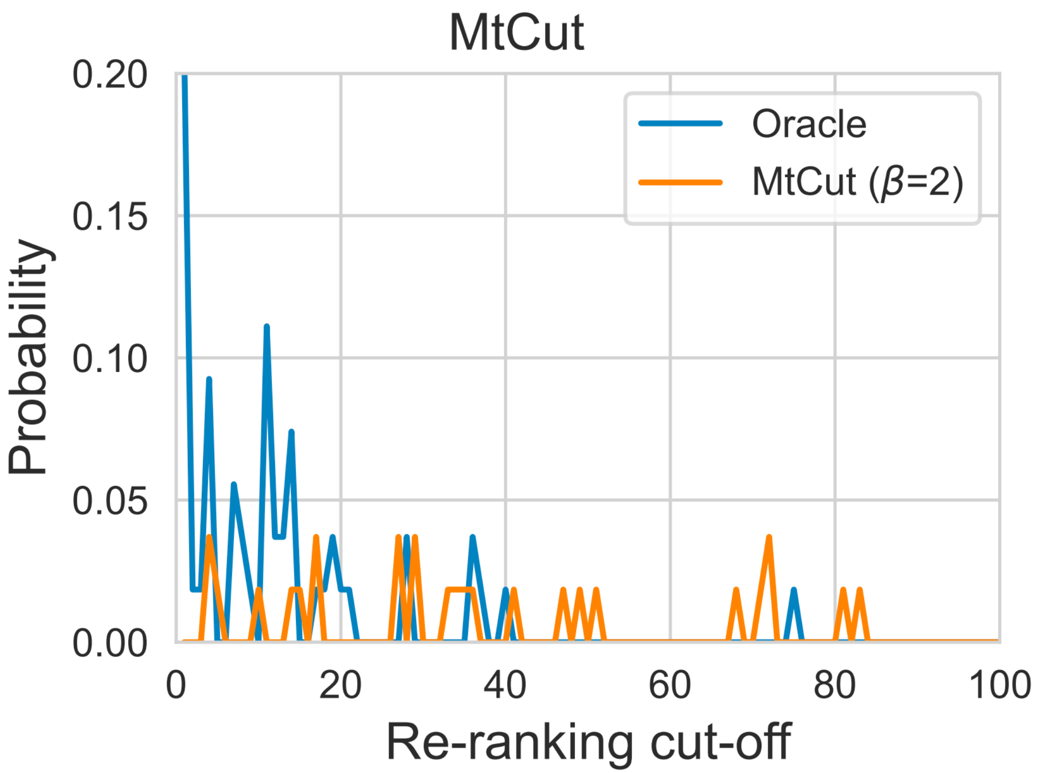
BM25–monoT5
TREC-DL 20

Experiments

- Error analysis for supervised RLT methods
  - They fail to predict a re-ranking cut-off of zero



Splade++–RankLLaMA
TREC-DL 20



RepLLaMA–RankLLaMA
TREC-DL 20

Takeaways

- The type of retriever makes a difference
  - With an effective retriever (e.g., SPLADE++/RepLLaMA)
    - A fixed re-ranking depth of **20** yields an excellent effectiveness/efficiency trade-off
    - A fixed depth>**20** does not significantly improve re-ranking quality

- The type of re-ranker (LLM or pre-trained LM-based) does not appear to influence the findings

- Supervised RLT methods need to improve their ability to predict “0”

Conclusion and Future Work

- Contributions
  - An empirical analysis in the context of LLM-based re-ranking, shows that
    - Effective query-specific re-ranking depths can improve re-ranking efficiency and effectiveness
  - We reproduce RLT methods in the context of LLM-based re-ranking
  - The data and code are open-source https://github.com/ChuanMeng/RLT4Reranking
- Future work
  - Explore RLT for pairwise and listwise LLM-based re-rankers
  - Develop new RLT methods for LLM-based re-ranking

Q & A

QR code for the repo

Outline

- ❑ Study 1: using LLMs as few-shot generative retriever [10 min]
- ❑ Study 2: using LLMs as relevance judgment and query performance predictor [10 min]
- ❑ Study 3: improve the efficiency of LLM-based re-rankers [15 min]
- ❑ **Conclusion** **[5 min]**

Conclusion

- Contributions
  - The opportunity to use LLMs as task solvers
    - Propose a Few-shot generative retrieval framework
  - The opportunity to use LLMs for evaluation
    - Fine-tune open-source LLMs to generate relevance judgments
    - A new QPP framework using LLM-based generated relevance judgments
  - The challenge of low efficiency in the context of LLM-based re-ranking
    - Predict query-specific re-ranking cut-offs

Personal website

Thank you!

c.meng@uva.nl

https://chuanmeng.github.io