Generating Synthetic Documents for Cross-Encoder Re-Rankers: A Comparative Study of ChatGPT and Human Experts

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, Suzan Verberne
Leiden University and University of Amsterdam

SIGIR TAIPEI | TAIWAN 2023

Study

- An investigation on the **usefulness of LLMs** in generating **training data** in a novel direction:
  - generating synthetic documents instead of synthetic queries
- We present the **ChatGPT-RetrievalQA dataset** for both full-ranking and re-ranking setups
  - With 24,322 queries, 26,882 responses generated by ChatGPT, and 58,546 human-generated responses
  - Queries from four different domains including Medicine, Finance, Reddit, and Wikipedia

Domain-level re-ranker effectiveness

Domain	Model	MAP	NDCG	Recall
Medicine	CE_{Human}	**.397**	**.419**	**.395**
	$CE_{ChatGPT}$	.379	.400	.377
Finance	CE_{Human}	**.257**	**.399**	**.251**
	$CE_{ChatGPT}$	.250	.368	.245
Reddit	CE_{Human}	**.323**	**.418**	**.543**
	$CE_{ChatGPT}$	.302	.391	.522
Wikipedia	CE_{Human}	.149	.152	.135
	$CE_{ChatGPT}$	**.163**	**.159**	**.144**

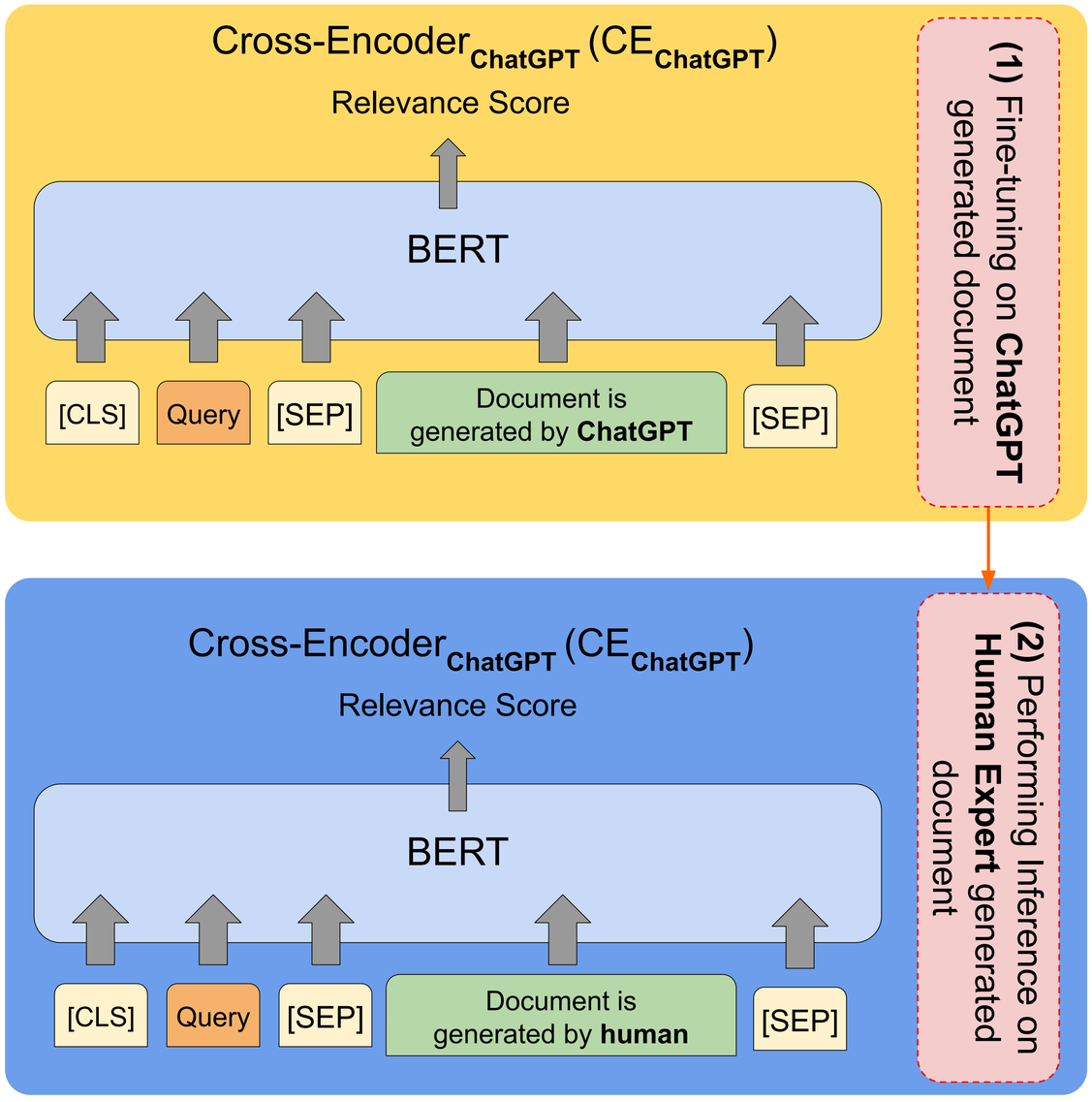
- Our results confirm that:
  - it is possible to train effective cross-encoder (CE) re-rankers by training them on ChatGPT-generated responses even for domain-specific queries.

BM25 on human VS. ChatGPT responses

Split	Source	MAP	NDCG	Recall
Test	Human	.143	.184	.520
	ChatGPT	.370	.396	.898

- BM25 is less effective on human-generated responses:
  - Indicating that human-generated responses are more challenging to match with queries
  - **Possible reason:** higher lexical overlap in the GPT output

Experimental setup



Main Results and Findings

- For **out-of-domain re-ranking:**
  - cross-encoder re-ranking models trained on LLM-generated responses are significantly more effective than those trained on human responses.
- For **in-domain re-ranking:**
  - The human-trained re-rankers outperform the LLM-trained re-rankers.

- **Conclusions:**
  - LLMs have **high potential** in generating training data for neural retrieval models and can be used to augment training data, especially in domains with smaller amounts of labeled data.
  - This work is particularly **advantageous** for **domain-specific tasks** where relying on LLM-generated output as a direct response to a user query can be risky.

Comparing the effectiveness of cross-encoder re-rankers fine-tuned on human and ChatGPT responses

Model	In-domain setting			Out-of-domain setting								
	ChatGPT-RetrievalQA (Ours)			TREC DL'19			TREC DL'20			MS MARCO DEV		
	MAP	NDCG	MRR	MAP	NDCG	MRR	NDCG	MRR	MAP	NDCG	MRR	MRR
BM25	.143	.184	.240	**.377**	.506	.858	.286	.480	.819	.195	.234	.187
$MiniLM_{Human}$	**.310**	**.384**	**.460**	.326	.451	.833	.269	.376	.913	.130	.155	.118
$MiniLM_{ChatGPT}$	.294	.362	.444	.342	**.510**	.903	**.344**	**.539**	**.978**	**.226**	**.267**	**.218**
$TinyBERT_{Human}$	.244	.310	.367	.294	.360	.741	.277	.364	.791	.128	.154	.116
$TinyBERT_{ChatGPT}$	.231	.291	.358	.328	.488	**.924**	.303	.460	.972	.194	.231	.185