

Generating Synthetic Documents for Cross-Encoder Re-**Rankers**:
**A comparative study of ChatGPT
versus human experts**

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, Suzan Verberne
Leiden Data Science Meetup
September 2023



Universiteit
Leiden

ChatGPT is not immune to mistakes or hallucinations



Treatments for Myeloproliferative neoplasms (MPN), include citations.



Myeloproliferative neoplasms (MPNs) are a group of rare blood disorders characterized by the overproduction of blood cells in the bone marrow. Treatment options for MPNs may vary depending on the specific type of MPN, the severity of the disease, and individual patient factors. Here are some common treatments for MPNs, along with citations to reputable sources:



3. **Cytoreductive Therapy:** Drugs like hydroxyurea, interferon-alpha, and ruxolitinib may be used to reduce blood cell production and manage symptoms in high-risk MPNs.

Citation: Vannucchi, A. M., Harrison, C. N., & Lucioni, M. (2018). Classical Philadelphia-negative myeloproliferative neoplasms: Diagnosis, treatment, and quality of life. Hematology, 2018(1), 129-139.

- Example of hallucination by ChatGPT
 - The above article does not exist in real-world and hallucinated by ChatGPT

ChatGPT is not
immune to
mistakes or
hallucinations

3. **Cytoreductive Therapy:** Drugs like hydroxyurea, interferon-alpha, and ruxolitinib may be used to reduce blood cell production and manage symptoms in high-risk MPNs.

Citation: Vannucchi, A. M., Harrison, C. N., & Lucioni, M. (2018). Classical Philadelphia-negative myeloproliferative neoplasms: Diagnosis, treatment, and quality of life. *Hematology*, 2018(1), 129-139.

- There are existing articles that share word overlap with the title of the hallucinated article.

Classical Philadelphia-negative myeloproliferative neoplasms: Diagnosis, treat



My profile

[HTML] Classical Philadelphia-negative myeloproliferative neoplasms: focus on mutations and JAK2 inhibitors

G Helbig - *Medical Oncology*, 2018 - Springer

Classical Philadelphia-negative myeloproliferative neoplasms (MPNs) encompass three main myeloid malignancies: polycythemia vera (PV), essential thrombocythemia (ET), and myelofibrosis (MF). Phenotype-driver mutations in Janus kinase 2 (JAK2), calreticulin (CALR), and myeloproliferative leukemia virus oncogene (MPL) genes are mutually exclusive and occur with a variable frequency. Driver mutations influence disease phenotype and prognosis. PV patients with JAK2 exon 14 mutation do not differ in number of ...

☆ Save Cite Cited by 22 Related articles All 9 versions

[HTML] springer.com

[HTML] Classical Philadelphia-negative myeloproliferative neoplasms (MPNs): A continuum of different disease entities

G Coltro, GG Loscocco, AM Vannucchi - *International Review of Cell and ...*, 2021 - Elsevier

Classical Philadelphia-negative myeloproliferative neoplasms (MPNs) are clonal hematopoietic stem cell-derived disorders characterized by uncontrolled proliferation of differentiated myeloid cells and close pathobiologic and clinical features. According to the 2016 World Health Organization (WHO) classification, MPNs include polycythemia vera (PV), essential thrombocythemia (ET), and primary myelofibrosis (PMF). The 2016 revision aimed in particular at strengthening the distinction between masked PV and JAK2-mutated ...

☆ Save Cite Cited by 12 Related articles All 5 versions

[HTML] sciencedirect.com

Search engines are
important because:
They do not
hallucinate!

☰ 🔍

📖 Scholar About 17.000 results (0,11 sec) YEAR ▾ ☰

Myeloproliferative neoplasms: a decade of discoveries and treatment advances [PDF] wiley.com
Full View
[A Tefferi](#) - [American journal of hematology](#), 2016 - [Wiley Online Library](#)
... **have** also been described in **MPN** and shown to provide additional prognostic information.
From the standpoint of **treatment**, ... basis for current **treatment** strategies in **MPN**. More recently, ...
☆ Save 📄 Cite Cited by 196 Related articles All 7 versions

A systematic review of antithrombotic treatment of venous thromboembolism in patients with myeloproliferative neoplasms [PDF] ashpublications.org
Full View
[EN Hamulyák](#), [JG Daams](#), [FWG Leebeek](#)... - [Blood](#) ..., 2021 - [ashpublications.org](#)
... **therapy**, with or without cytoreduction, in **MPN** patients with a history of VTE. A total of 5675 unique **citations** ... the optimal antithrombotic **treatment** of VTE in patients with **MPN** is based on ...
☆ Save 📄 Cite Cited by 35 Related articles All 10 versions

Risk of hepatitis B virus reactivation following ruxolitinib treatment in patients with myeloproliferative neoplasms [HTML] tandfonline.com
[M Duan](#), [XX Cao](#), [L Chang](#), [D Zhou](#) - [Hematology](#), 2021 - [Taylor & Francis](#)
... with **myeloproliferative neoplasms (MPN)** and chronic or resolved HBV infection during and after ruxolitinib **treatment** ... **treatment**, which is consistent with previous findings [**Citation**11,...
☆ Save 📄 Cite Cited by 6 Related articles All 3 versions

Impact of gene mutations on treatment response and prognosis of acute myeloid [PDF] wiley.com

- Search engines find the most relevant documents
 - Document are written by human and source of each retrieved document is transparent.

Search engines are
important because:
They do not
hallucinate!

☰ Treatments for Myeloproliferative neoplasms (MPN), include citations. 🔍

📖 Scholar About 17.000 results (0,11 sec) YEAR ▾ ☰

Myeloproliferative neoplasms: a decade of discoveries and treatment advances [PDF] wiley.com
Full View
[A Tefferi](#) - [American journal of hematology](#), 2016 - Wiley Online Library
... **have** also been described in **MPN** and shown to provide additional prognostic information.
From the standpoint of **treatment**, ... basis for current **treatment** strategies in **MPN**. More recently, ...
☆ Save 📄 Cite Cited by 196 Related articles All 7 versions

A systematic review of antithrombotic treatment of venous thromboembolism in patients with myeloproliferative neoplasms [PDF] ashpublications.org
Full View
[EN Hamulyák](#), [JG Daams](#), [FWG Leebeek](#)... - [Blood](#) ..., 2021 - ashpublications.org
... **therapy**, with or without cytoreduction, in **MPN** patients with a history of VTE. A total of 5675 unique **citations** ... the optimal antithrombotic **treatment** of VTE in patients with **MPN** is based on ...
☆ Save 📄 Cite Cited by 35 Related articles All 10 versions

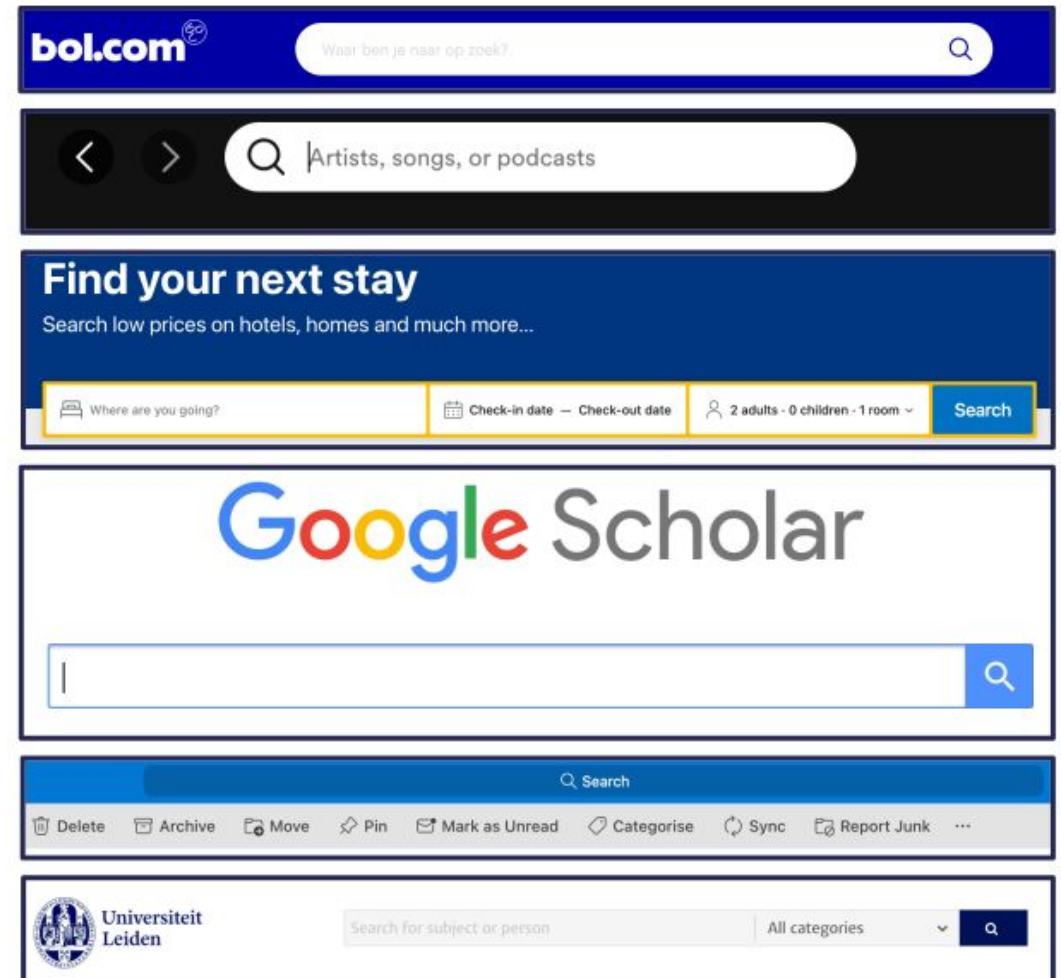
Risk of hepatitis B virus reactivation following ruxolitinib treatment in patients with myeloproliferative neoplasms [HTML] tandfonline.com
[M Duan](#), [XX Cao](#), [L Chang](#), [D Zhou](#) - [Hematology](#), 2021 - Taylor & Francis
... with **myeloproliferative neoplasms (MPN)** and chronic or resolved HBV infection during and after ruxolitinib **treatment** ... **treatment**, which is consistent with previous findings [**Citation**11,...
☆ Save 📄 Cite Cited by 6 Related articles All 3 versions

Impact of gene mutations on treatment response and prognosis of acute myeloid [PDF] wiley.com

- Search engines find the most relevant documents
- Main component of search engines: **retrieval models** (rankers)
- But, search engines need to be optimized/trained on relatively large amount of **training data**

Retrieval objective

- **Retrieval objective:**
 - Given an unseen question
 - Providing a list of most relevant documents.
- Example of search tasks that are dependent on retrieval models:



Summary:

- The **lack of transparency** in the source of information generated by ChatGPT can be a bigger concern in domains such as law, **medicine**, and science, where
 - accountability and trustworthiness are critical
- **Ranking models provide more reliability:**
 - as opposed to generative models, they retrieve information
 - source of each retrieved item is **transparent**.
- Therefore, document retrieval – even when generative LLMs are available – remains an **important application**, especially in mentioned domains.
- **However**, training document retrieval models requires large amount of training data that is labeled by human which is a costly task.
- **Our main question in this paper:** would it be possible to use ChatGPT to build retrieval models?

Domains considered in this research

- We use four specific domains:
 - Medicine
 - Finance
 - Reddit
 - Wikipedia
- and build ChatGPT-RetrievalQA dataset
 - based on an existing dataset, human ChatGPT Comparison Corpus (HC3).



Scan the QR Code to
check out the dataset

Example of ChatGPT-RetrievalQA dataset in which each question is answered by a human expert and ChatGPT:

Question: Is chest pain related to intake of clindamycin and oxycodone? Hi, I was recently released from the hospital after a hand surgery and they provided me with Clindamycin 300mg and Oxycodone Acetaminophens. I've taken this combination 3 times now and my chest feels really tight. Is there reason for me to worry?

Human Expert:

The use of Clindamycin can cause stomach pain or a hyperacidity of the stomach.
So, I recommend using a medication to lower the acidity production such as Omeprazole daily.
I also suggest using Maalox three times a day and avoid food that can trigger the symptoms such as spicy food.
Hope I have answered your query. Let me know if I can assist you further.

Take care.
Regards,
Dr. Dorina Gurabardhi,
General & Family Physician.

ChatGPT:

It is possible that chest pain could be related to the use of Clindamycin and Oxycodone.
However, it is also important to consider other potential causes of chest pain, such as heart problems or other underlying health conditions. It is important to speak with a healthcare provider if you are experiencing chest pain, as this could be a serious issue that requires medical attention. In the meantime, you should follow the instructions of your healthcare provider and report any adverse effects to them. If you are experiencing difficulty breathing or severe chest pain, you should seek immediate medical attention.

Goal:

- Training retrievers using ChatGPT response

Question:

- What is the effectiveness of rankers trained on ChatGPT-generated responses compare to those trained on human-generated responses?

Human Expert:

The use of Clindamycin can cause stomach pain or a hyperacidity of the stomach.

So, I recommend using a medication to lower the acidity production such as Omeprazole daily.

I also suggest **using Maalox three times** a day and **avoid** food that can trigger the symptoms such as **spicy food**.

Hope I have answered your query. Let me know if I can assist you further.

Take care.

Regards,

Dr. Dorina Gurabardhi,
General & Family Physician.

ChatGPT:

It is possible that chest pain could be related to the use of Clindamycin and Oxycodone.

However, it is also important to consider other potential causes of chest pain, such as heart problems or other underlying health conditions. **It is important to speak with a healthcare provider** if you are experiencing chest pain, as this could be a serious issue that requires medical attention. In the meantime, you should follow the instructions of your healthcare provider and report any adverse effects to them. If you are experiencing difficulty breathing or severe chest pain, you should **seek immediate medical attention**.

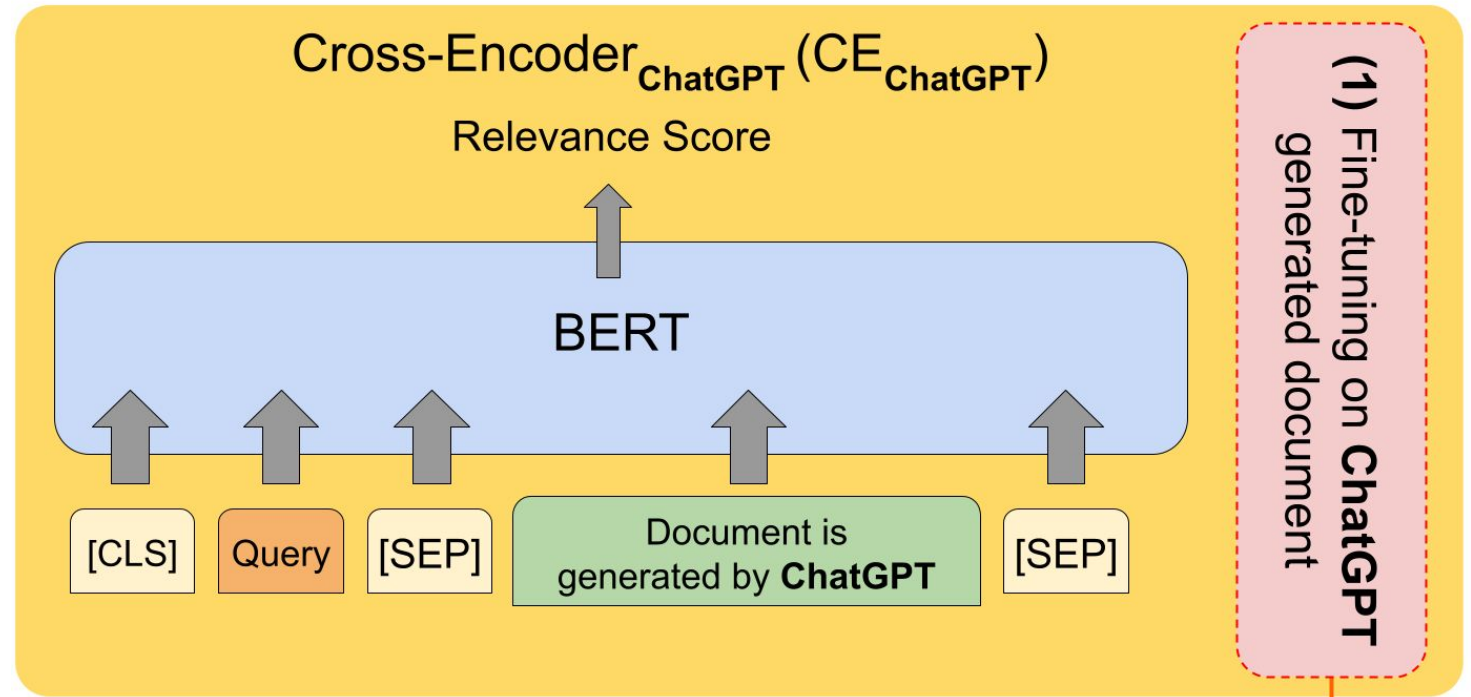
Dataset statistics

Table 1: Statistic on the size of Train, Validation, and Test sets across domains for evaluation of cross-Encoders.

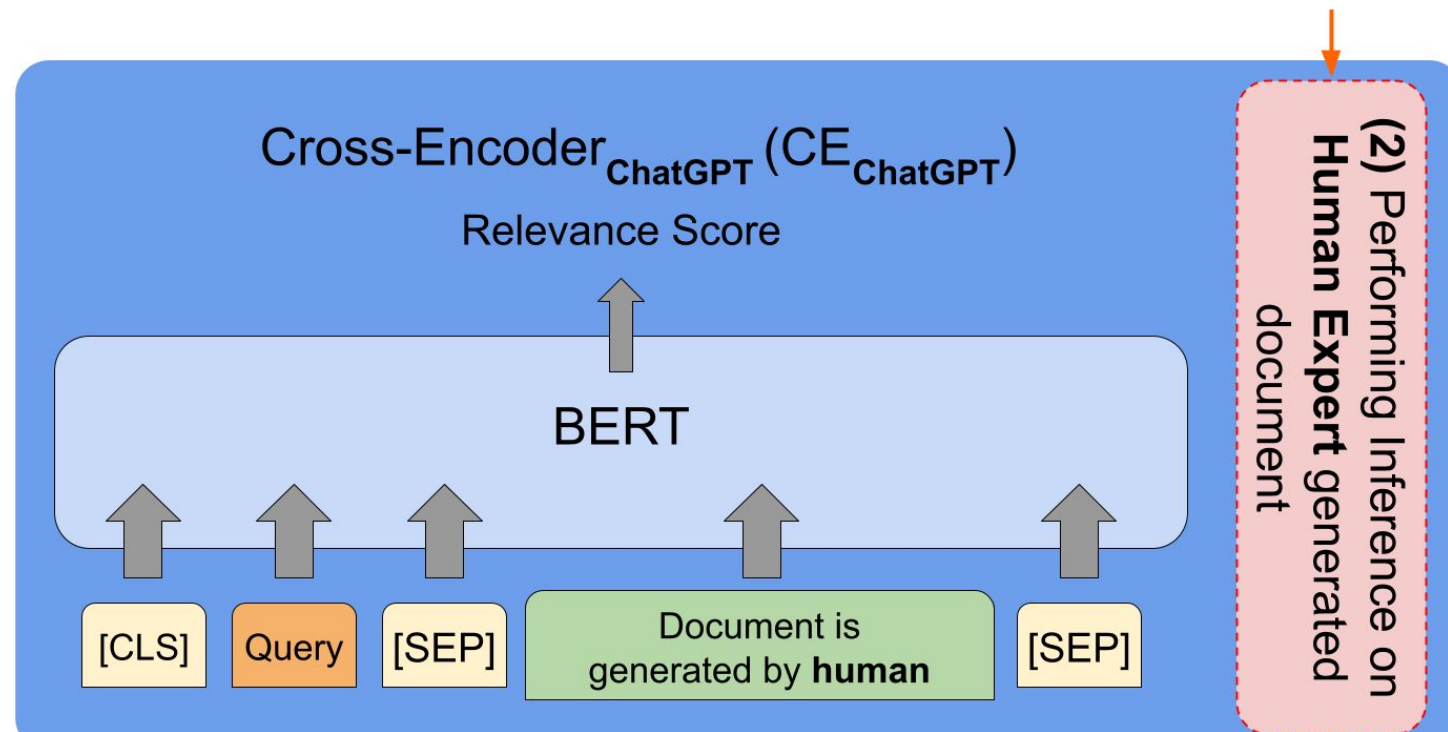
Domain	# of queries		
	Train set	Validation set	Test set
All	16788	606	6928
Medicine: Meddialog [7]	862	31	355
Finance: FiQA [24]	2715	98	1120
Reddit: ELI5 [13]	11809	427	4876
Wikipedia: openQA [40]	820	29	338
Wikipedia: csai [15]	582	21	239

Experimental setup

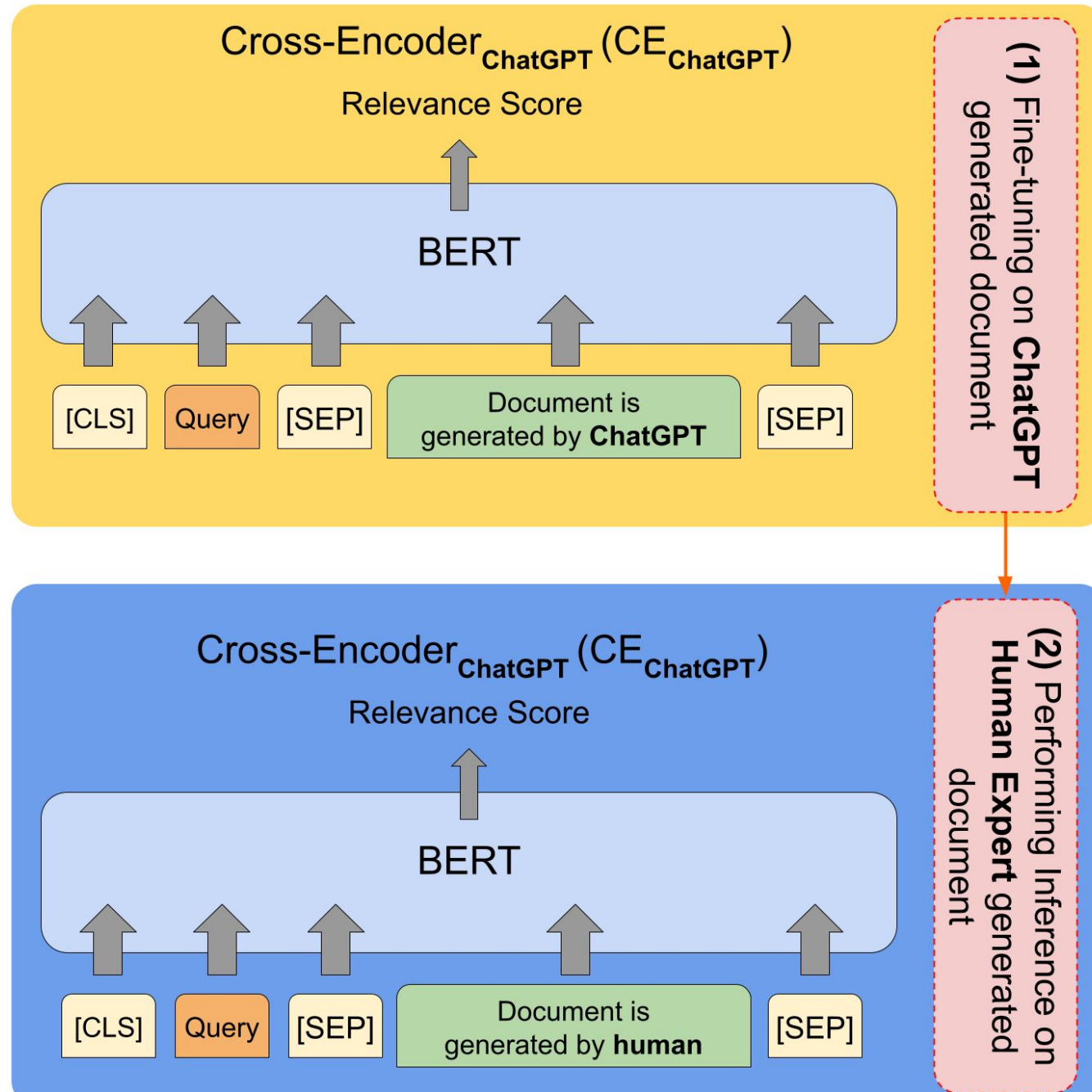
- We train one of the current strong retrieval models, called **cross-encoders**.
- Next, for **evaluation**, we use human responses.

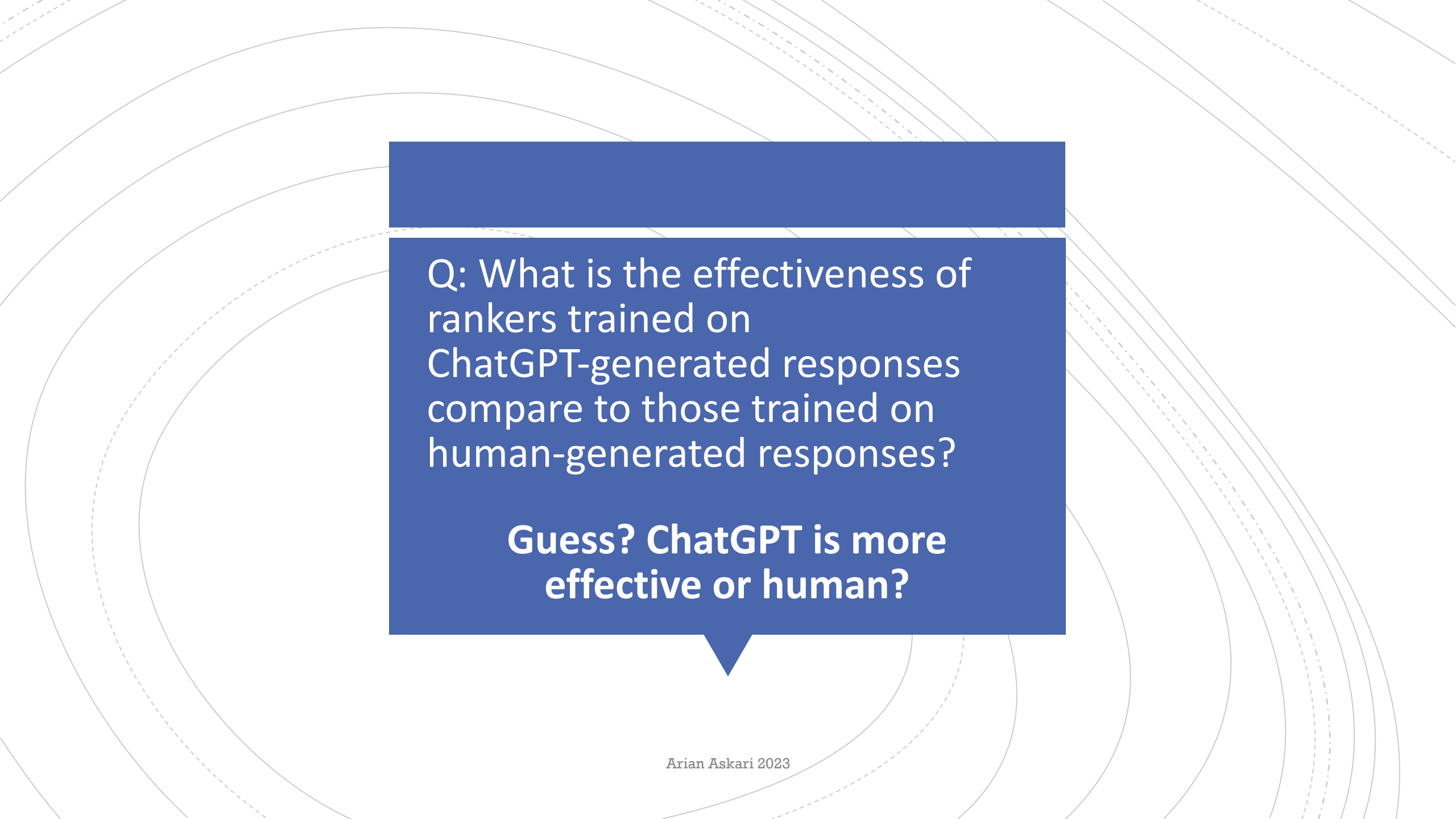


Experimental setup



Experimental setup



The background features a series of concentric circles in light gray, some solid and some dashed, creating a ripple effect. In the center, there is a blue speech bubble with a white border. The text is contained within this bubble.

Q: What is the effectiveness of
rankers trained on
ChatGPT-generated responses
compare to those trained on
human-generated responses?

**Guess? ChatGPT is more
effective or human?**

What is the effectiveness of rankers trained on ChatGPT-generated responses compare to those trained on human-generated responses?

- Results on **Medicine domain** (the higher is the better/more effective)
 - human-trained models are slightly more effective
 - This is while making human labeled is a much more costly task usually

Domain	Model	MAP@1K	NDCG@10	MRR@10
Medicine	human-trained	.397	.419	.395
	ChatGPT-trained	.379	.400	.377

Results on **Wikipedia domain** (the higher is the better/more effective)

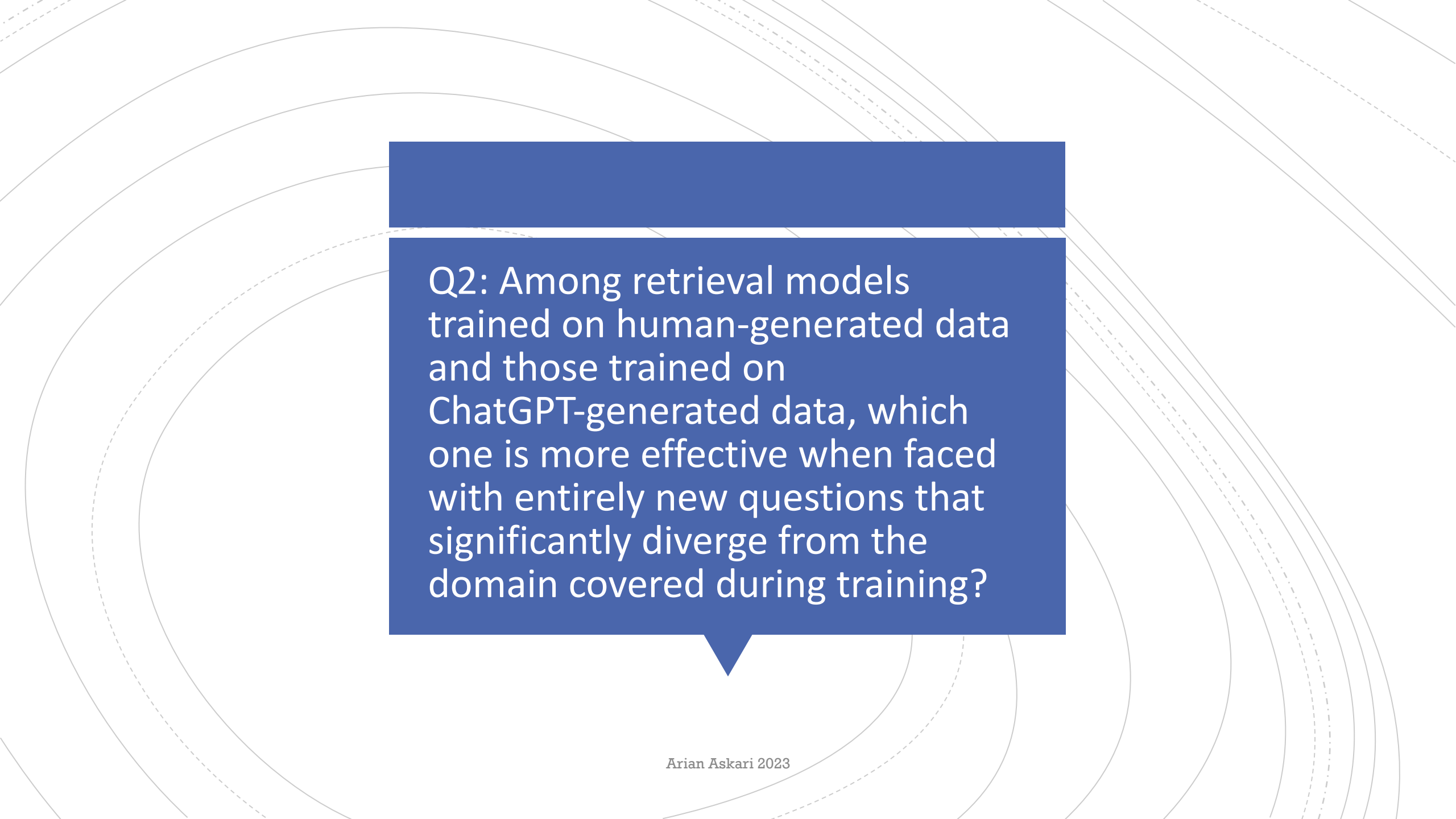
- human-trained models are less effective!
- This could be partly because that ChatGPT has wikipedia database in its training data
 - Therefore, it is able to produce high quality response for this questions

Domain	Model	MAP@1K	NDCG@10	MRR@10
Wikipedia	human-trained	.149	.152	.135
	ChatGPT-trained	.163	.159	.144

Results on **all domains** (the higher is the better/more effective)

- human-trained models are on average slightly more effective.

Domain	Model	MAP@1K	NDCG@10	MRR@10
All	human-trained	.310	.384	.460
	ChatGPT-trained	.294	.362	.444
Finance	human-trained	.257	.399	.251
	ChatGPT-trained	.250	.368	.245
Medicine	human-trained	.397	.419	.395
	ChatGPT-trained	.379	.400	.377
Wikipedia	human-trained	.149	.152	.135
	ChatGPT-trained	.163	.159	.144

The background features a series of concentric circles in light gray, some solid and some dashed, creating a ripple effect. A solid blue speech bubble is centered on the page, containing white text. The bubble has a rectangular body and a triangular tail pointing downwards.

Q2: Among retrieval models trained on human-generated data and those trained on ChatGPT-generated data, which one is more effective when faced with entirely new questions that significantly diverge from the domain covered during training?

- Results on TREC DL'20 dataset:
 - ChatGPT-trained models are statistically significantly more effective than human-trained models for entirely new question.
 - ChatGPT-trained models can generalize better in our experiments

Model	Training source	MAP@1K	NDCG@10	MRR@10
Unsupervised model (BM25)	—	.286	.480	.819
MiniLM	human-trained	.269	.376	.913
	ChatGPT-trained	.344	.539	.978
TinyBERT	human-trained	.277	.364	.791
	ChatGPT-trained	.303	.460	.377

Limitations:

1. Other aspects of evaluation have not been investigated in this paper, specifically:
 - the **effect of biased** information in the generated documents on biases in the document ranking.
2. Another problem is that the **factuality** of the **ChatGPT output cannot be guaranteed**.
 - Even though, factually incorrect information in the generated data (as a result of ChatGPT's hallucination) is not likely to be harmful in the retrieval context, because only information that is truly contained in the document collection can be retrieved by a retrieval model in inference.
3. We do not systematically study and quantify the **effect of hallucinated data on the performance** of the ranker.

Takeaways

- For **out-of-domain ranking**: retrieval models trained on LLM-generated responses are significantly more effective than those trained on human responses.
- For **in-domain ranking**: human-trained re-retrieval models outperform the LLM-trained re-rankers.
- 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.
- Our work [1,2] is particularly **advantageous for domain-specific** tasks where relying on LLM-generated output as a direct response to a user query can be risky.
- There are also works that focus on **generating queries** [3] and **relevance assessments** [4].

[1] Askari et al (CIKM 2023). A Test Collection of Synthetic Documents for Training Rankers: ChatGPT vs. Human Experts

[2] Askari et al (GenIR SIGIR 2023). Generating Synthetic Documents for Cross-Encoder Re-Rankers: A Comparative Study of ChatGPT and Human Experts

[3] Jeronymo et al. (arXiv 2023) InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval

[4] Faggioli et al. (ICTIR 2023). Perspectives on Large Language Models for Relevance Judgment

Takeaways

- For **out-of-domain ranking**: retrieval models trained on LLM-generated responses are significantly more effective than those trained on human responses.
- For **in-domain ranking**: human-trained re-retrieval models outperform the LLM-trained re-rankers.
- 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.
- Our work [1,2] is particularly **advantageous for domain-specific** tasks where relying on LLM-generated output as a direct response to a user query can be risky.
- There are also works that focus on **generating queries** [3] and **relevance assessments** [4].

Thank you!



Scan the QR Code to
check out the dataset