Content Accuracy of the Gpt4all-Falcon Model

Patrick Hanfstingl | Florian Zanotti | Jakob Zenz

Group 3

23.01.2024

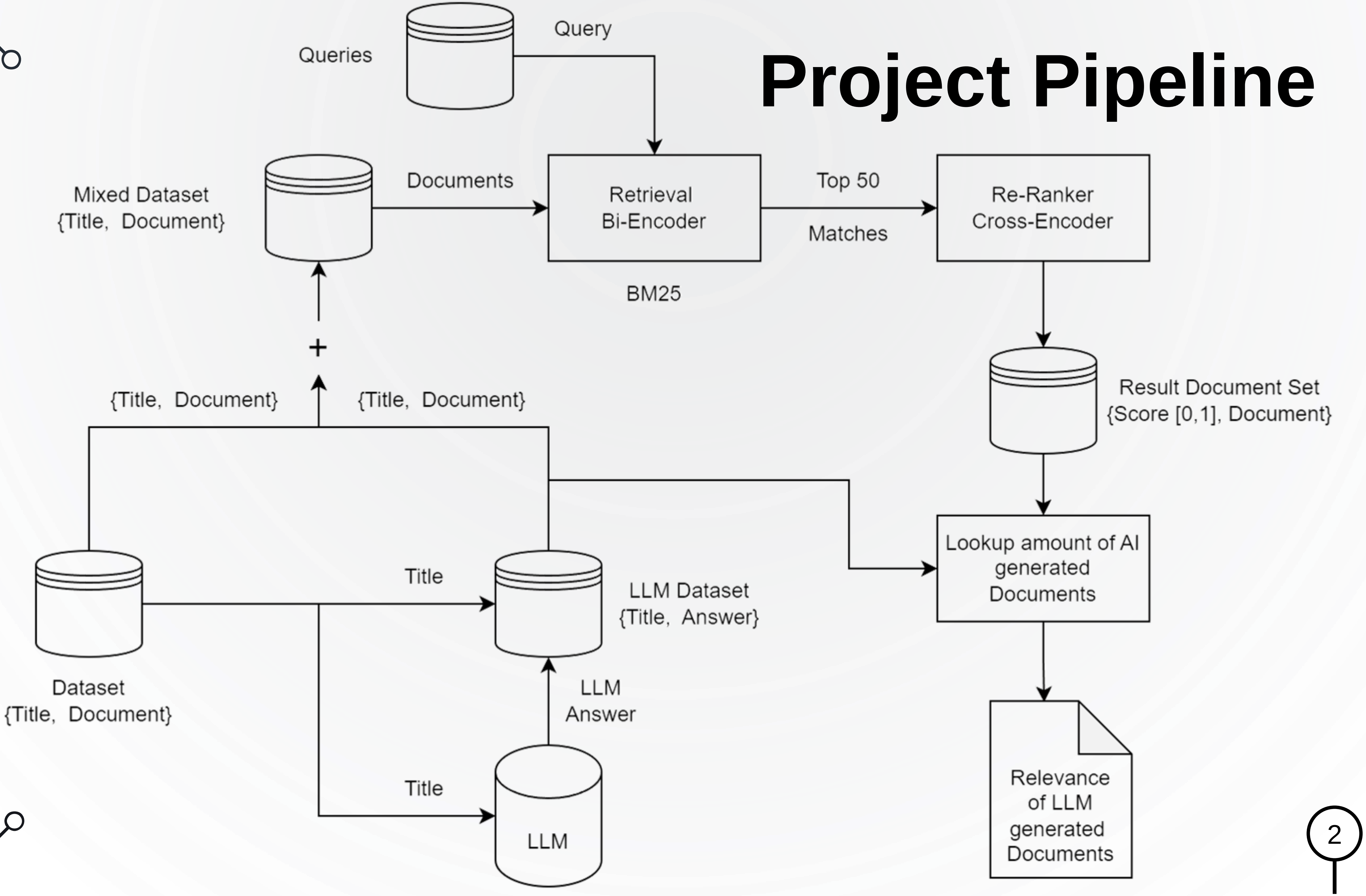
How do LLMs rank?

- Rise of Large Language Models (ChatGPT etc.)
- Ranking of generated content
- Relevancy of the content
- Validation of our pipeline

[1]

[1] https://www.linkedin.com/pulse/llm-revolution-how-ai-language-models-transforming-lives-ahmed-jawed

Project Pipeline



Choosing the Transformer

Requirements:

- Capable of relevance classification
- Performs well on our validation data

MonoBert [1]:

- MonoBert adapts Bert for relevance classification
- Open source and well known
- Easy to use

[1] https://huggingface.co/castorini/monobert-large-msmarco

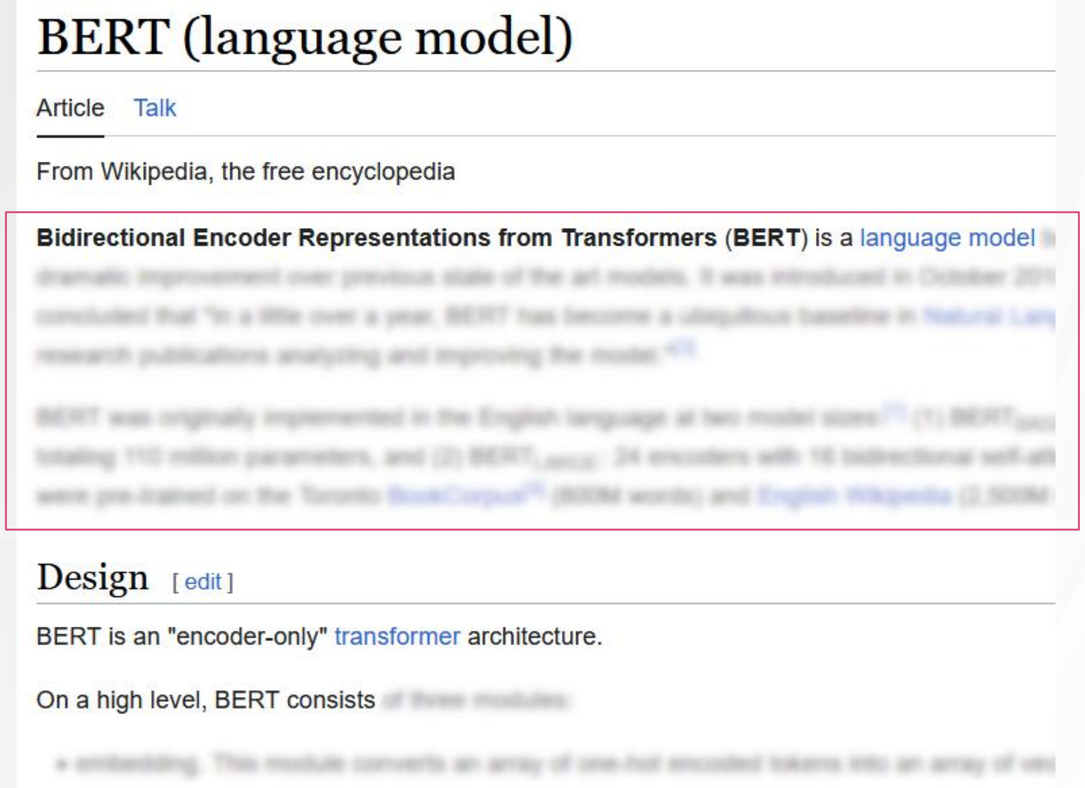
Choosing the Dataset

Requirements:

- Collection of general, broad knowledge
- Contains queries
- Not too large

Wikipedia summary dataset [1]:

- Contains summaries of Wikipedia pages
- Titles as queries
- subset “sport”



BERT (language model)

Article Talk

From Wikipedia, the free encyclopedia

Bidirectional Encoder Representations from Transformers (**BERT**) is a language model
dramatic improvement over previous state of the art models. It was introduced in October 201
concluded that "in a little over a year, BERT has become a ubiquitous baseline in Natural Lang
research publications analyzing and improving the model."[4]

BERT was originally implemented in the English language at two model sizes:[1] (1) $\text{BERT}_{\text{BASE}}$
totaling 110 million parameters, and (2) $\text{BERT}_{\text{LARGE}}$: 24 encoders with 16 bidirectional self-att
were pre-trained on the Toronto BookCorpus[5] (800M words) and English Wikipedia (2,500M

Design [edit]

BERT is an "encoder-only" transformer architecture.

On a high level, BERT consists of three modules:

- embedding. This module converts an array of one-hot encoded tokens into an array of vec

Wikipedia entry of “Bert“ [2]

[1] https://github.com/tscheepers/Wikipedia-Summary-Dataset

[2] https://en.wikipedia.org/wiki/BERT_(language_model)

LLM (GPT4-ALL)

- Locally run, train and deploy LLMs
- Open source
- Maintained by Noimic AI
- GPT4All-Falcon: very **fast**, good **quality**
- Processed titles from dataset to get query
- Problems with chat session:
  - Optimized it to take less time
  - Get right format as response

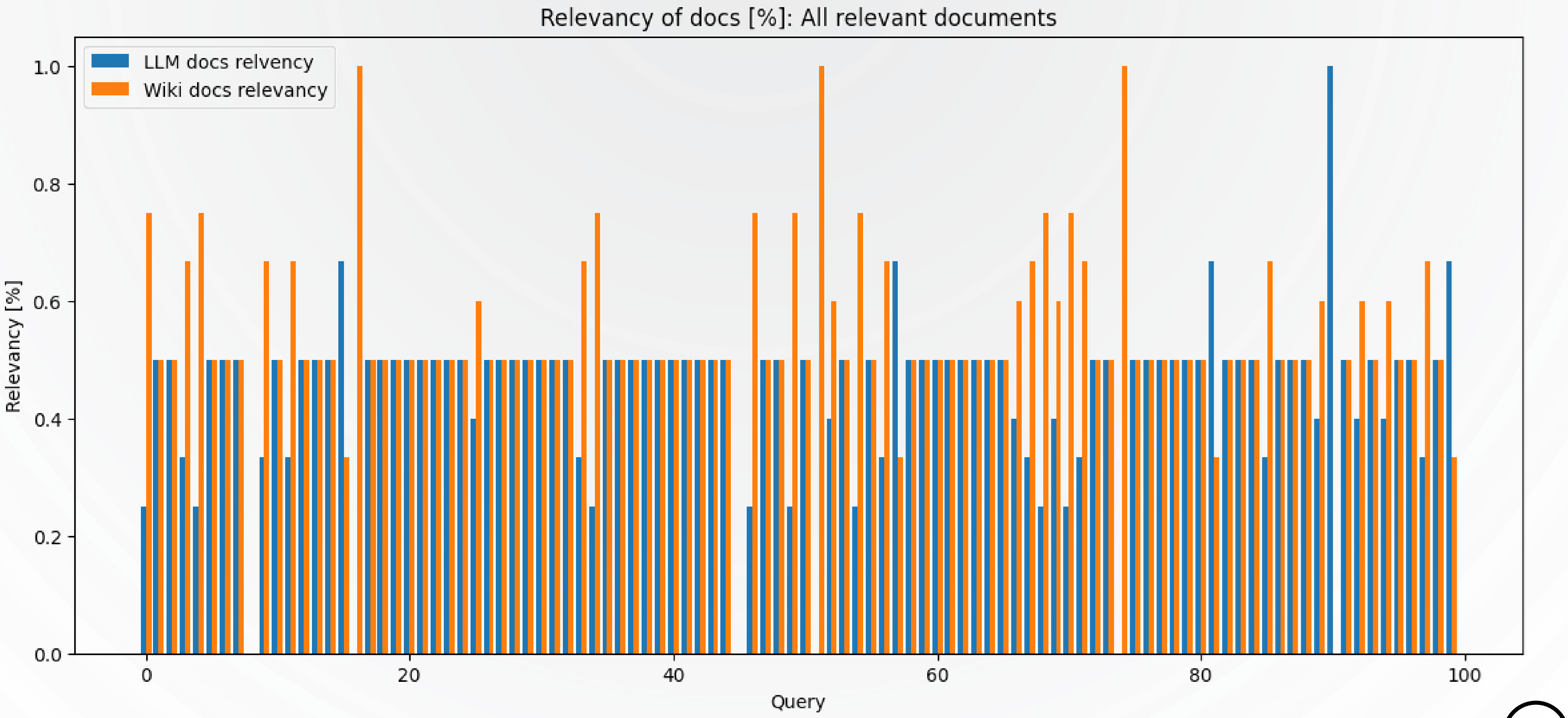
Validation of our pipeline

- Important that our model is sound
- nDCG@k and f1@k
- Dataset Wikipedia articles [1]
- Subset of the same topic

	BM25	BM25 + MonoBert
f1@k:	0.824	0.824
nDCG@k:	0.9816	0.9886

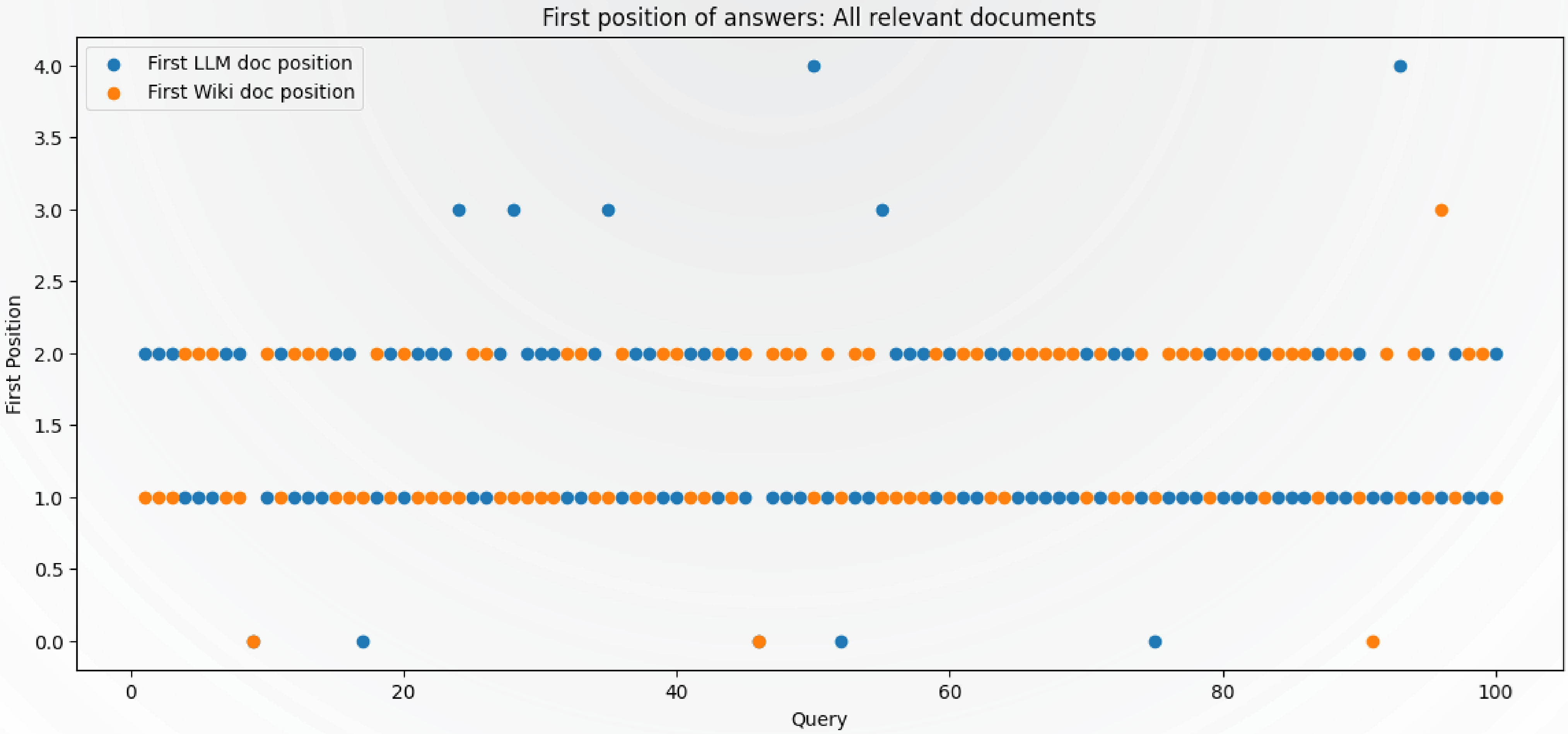
[1] https://ir-datasets.com/wikir.html

Amount of LLM documents [%]



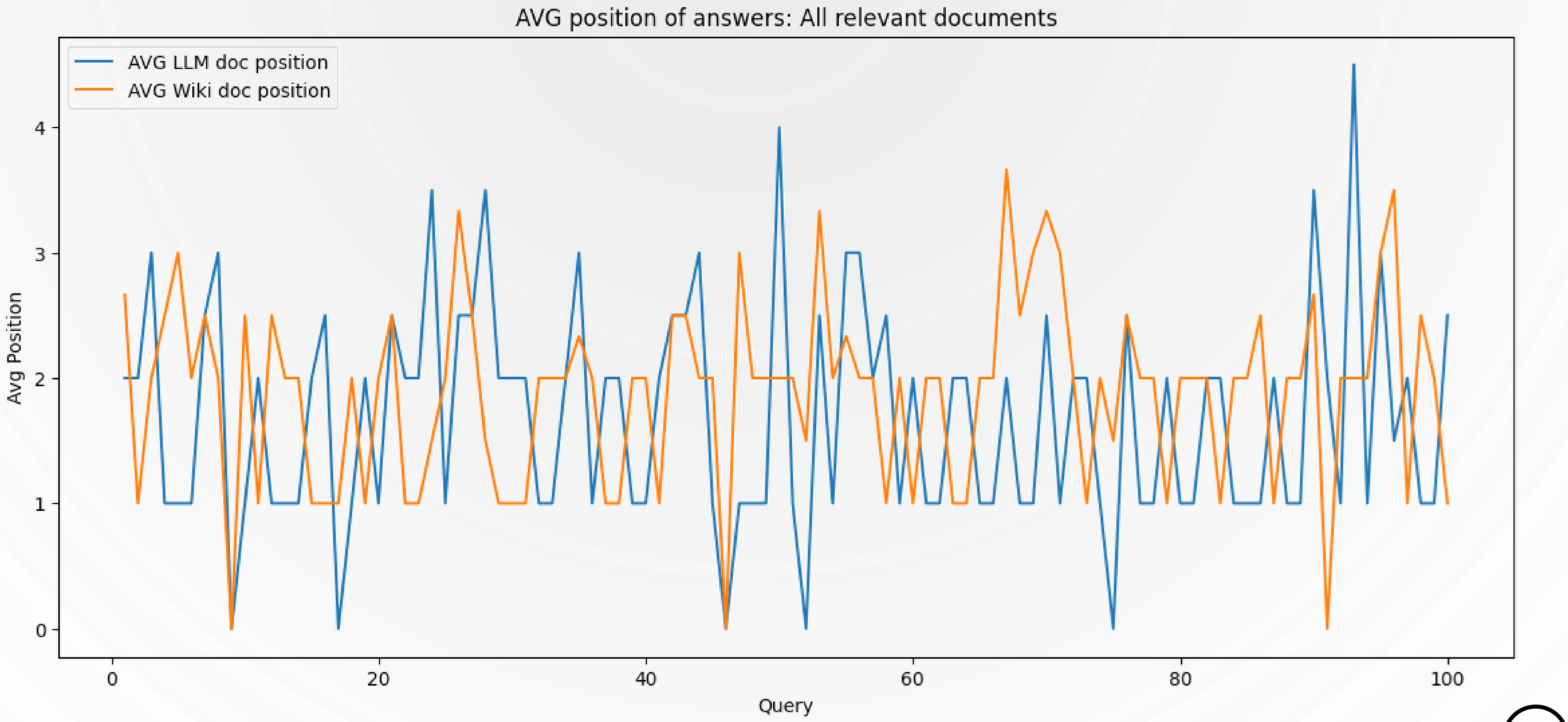
LLM ≈ 45%; Wiki ≈ 55%

Position of first LLM document



LLM ≈ 1.47; Wiki ≈ 1.48

Average position



LLM ≈ 1.69; Wiki ≈ 1.88

Results:

- Valuable information
- Ranks as high as Wiki
- Some outliers

- Details: README [1]

[1] https://github.com/zanflo/AdvancedInformationRetrieval

Conclusion

Gpt4all-falcon is content accurate

… out of the box

... on general sport topics

Content is nearly as accurate as (human written) facts

Future Improvements

- Select a different dataset for testing
  - Different query – stronger semantics
- Get more resources
  - use larger datasets to exclude outliers
  - test the model on general knowledge

Questions?

https://github.com/zanflo/AdvancedInformationRetrieval