Harnessing Information Retrieval Techniques in Retrieval-Augmented Generation

Lixiang Li, Wendy Jiang, Chen Peng, Jie Zheng
{li4256, jiang794, peng326, zheng795} @purdue.edu

PURDUE UNIVERSITY | Department of Computer Science

Outline

- Introduction 4m
- Background: 4m
    - Vector Database
    - Marginalization and Embedding
    - Information Retrieval
    - Large Language Model (LLM)
- Methodology 2m
    - Re-ranking
- Experiment Setup 3m
    - Datasets, Metrics, benchmarks
- Experimental Results and Insights 3m

Background

- Retrieval-Augmented Generation:
  - A technique that combines the strengths of retrieval-based and generation-based models.

- Improve Factual Accuracy
- Provide context information
- Provide up to date information

Background

- Limitation:
  - LLM relies on highly relevant and specific information.
  - But traditional methods don’t provide relevant and effective information
  - Traditional methods provides long top k documents which are too long to be effective
  - Truncation methods are not reliable enough

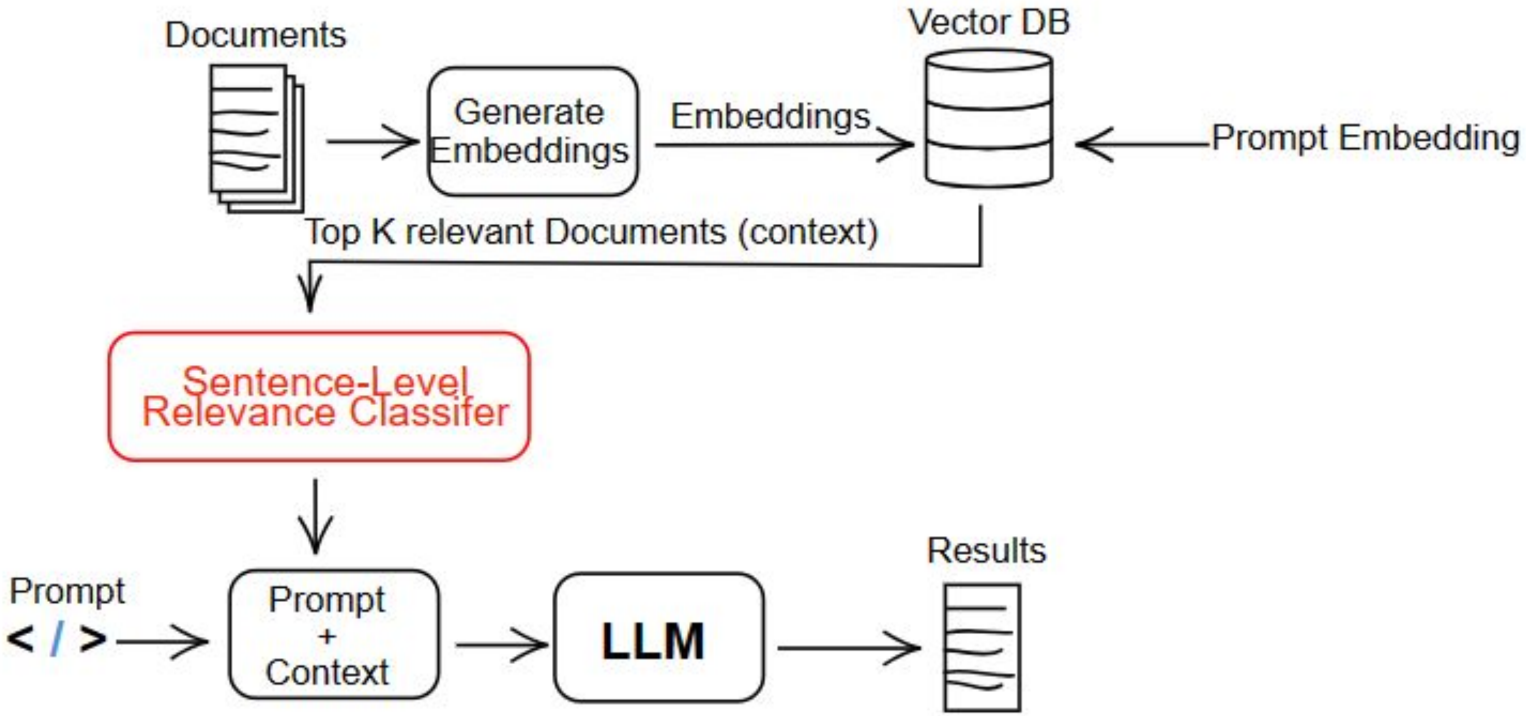
Research Question

- How to design a model to help RAG to get the relevant and important information?

- We propose a model to identify most relevant information.

Proposed Method

•

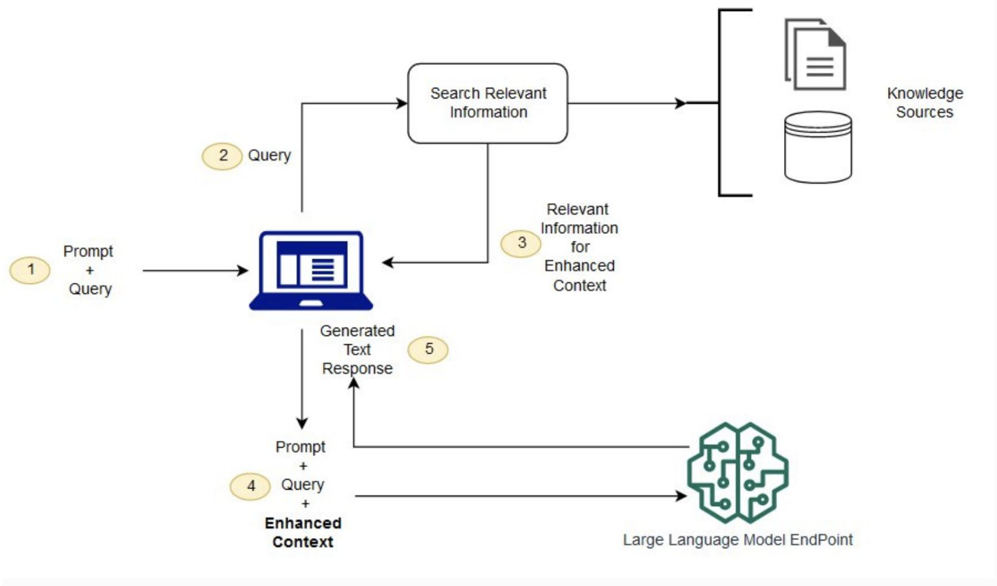


Related Work

- Retrieval Methods in RAG
  - Can’t perform well for long documents
- Truncation Methods in RAG
  - Not reliable since it’s hard to determine the truncation cutoff.

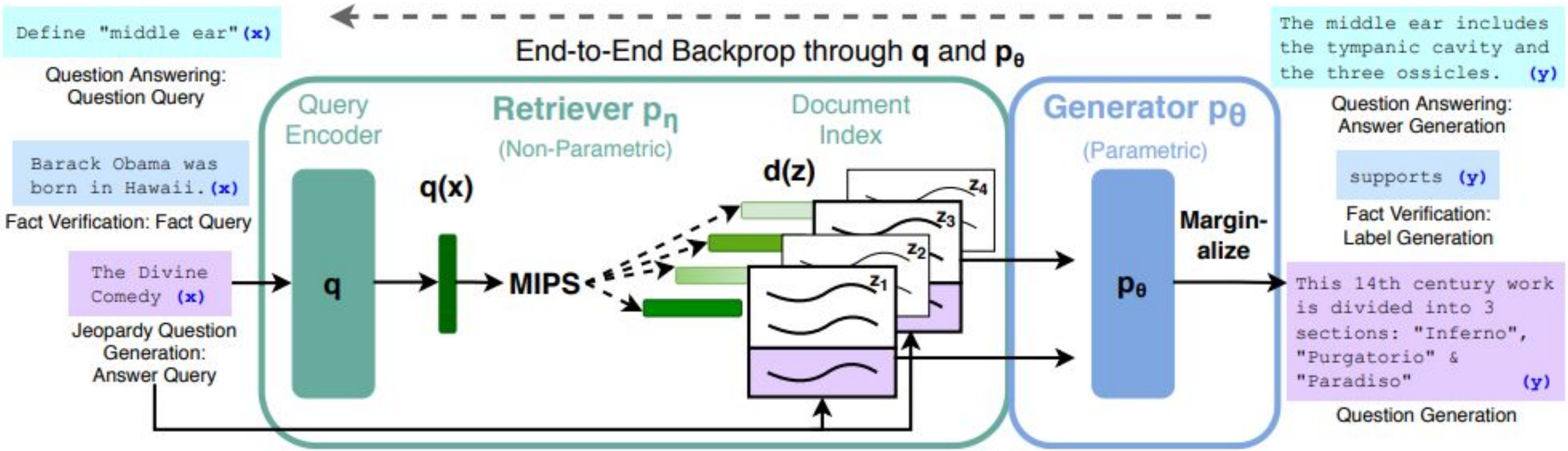
Overview

- Benefits of RAG: improved model performance, information retrieval, and cost efficiency.
- Challenges: potential biases in datasets, computational complexity.



https://aws.amazon.com/what-is/retrieval-augmented-generation/

RAG



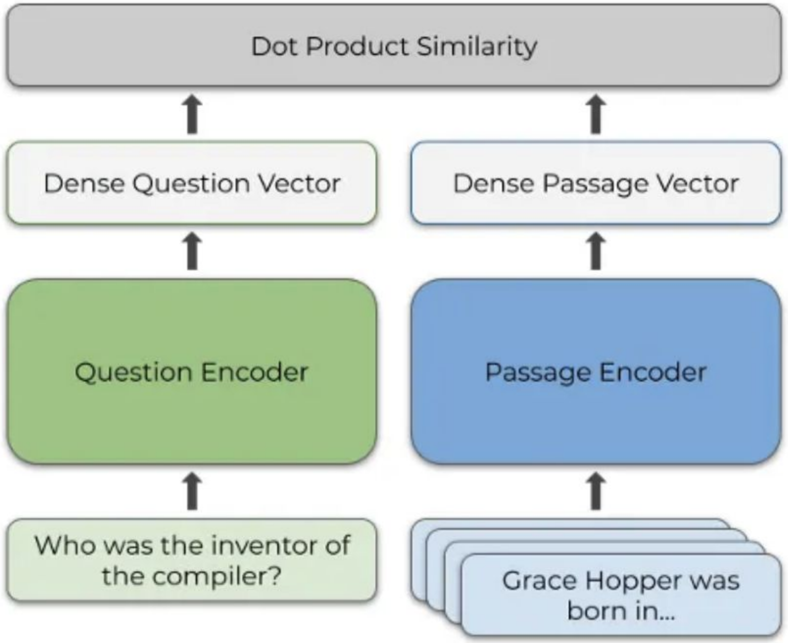
arXiv:2005.11401

RAG Sequence and RAG Token

$$p_{\text{RAG-Sequence}}(y|x) \approx \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) = \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_i^N p_\theta(y_i|x,z,y_{1:i-1})$$

$$p_{\text{RAG-Token}}(y|x) \approx \prod_i^N \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y_i|x,z,y_{1:i-1})$$

RAG Retriever



Dense Passage Retrieval (DPR) architecture.

$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right)$$

$$\mathbf{d}(z) = \mathrm{BERT}_d(z), \quad \mathbf{q}(x) = \mathrm{BERT}_q(x)$$

https://shamanesiri.medium.com/how-to-finetune-the-entire-rag-architecture-including-dpr-retriever-4b4385322552
https://arxiv.org/pdf/2005.11401.pdf

Embedding Models

Overall | Bitext Mining | Classification | Clustering | Pair Classification | Reranking | Retrieval | STS | Summarization

English | Chinese | French | Polish

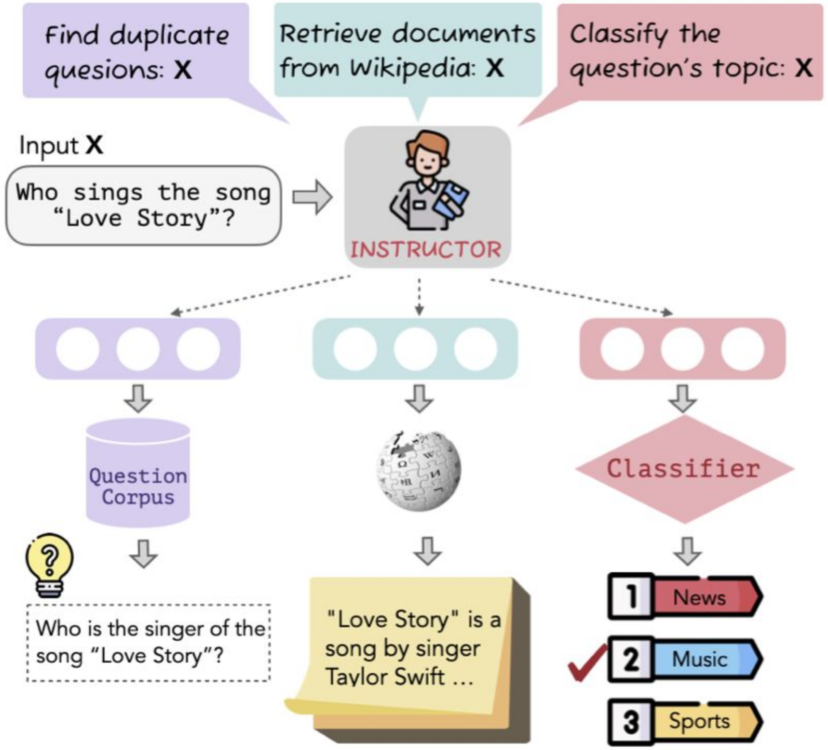
Overall MTEB English leaderboard 🔮

- **Metric:** Various, refer to task tabs
- **Languages:** English

Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)
1	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67
2	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13	79.25	52.42
3	GritLM-7B	7242	26.98	4096	32768	66.76	79.46	50.61
4	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63	78.47	50.26
5	google-gecko.text-embedding-p	1200	4.47	768	2048	66.31	81.17	47.48
6	GritLM-8x7B	46703	173.98	4096	32768	65.66	78.53	50.14
40	instructor-large	335	1.25	768	512	61.59	73.86	45.29

https://huggingface.co/spaces/mteb/leaderboard

Instructor Model




```
import pickle
with open('NQDataset_with_ContentEmbeddings.pkl', 'wb') as f:
  pickle.dump(df_copy, f)
```

```
with open('NQDataset_with_ContentEmbeddings.pkl', 'rb') as f:
  loaded_df = pickle.load(f)
```

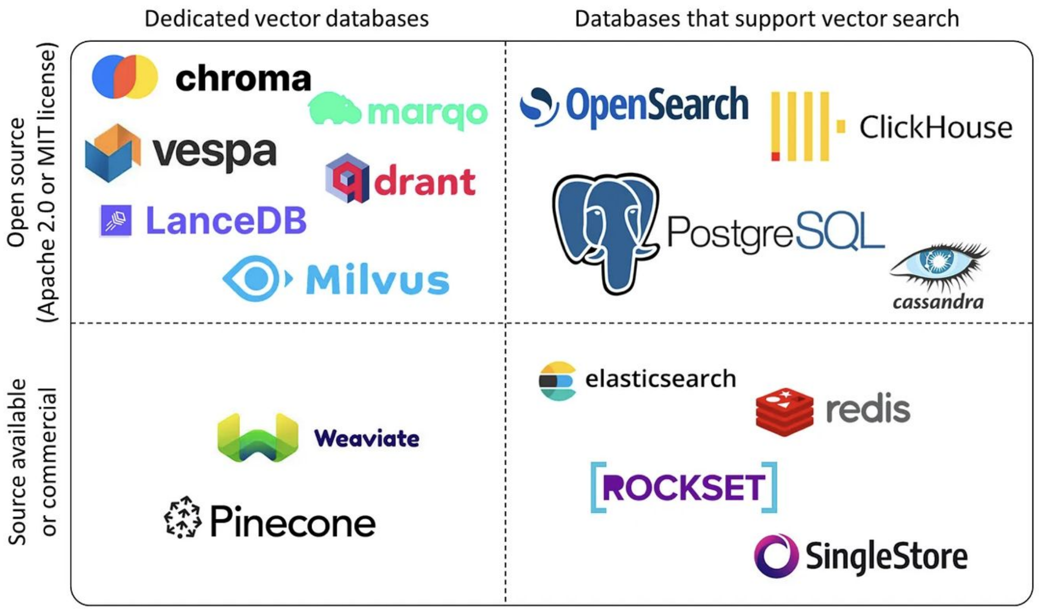
```
loaded_df.iloc[0]
```

```
Unnamed: 0                                                            0
query                                  wolf of wall street number of f words
long_answer           Film    Year    Fuck count    Minutes    Uses / mi...
short_answer                                                        569
title                 List of films that most frequently use the wor...
bert_title            list of films that most frequently use the wor...
abstract              The use of profanity in films has always been ...
content               This is a list of non-pornographic , English l...
url                   https://en.wikipedia.org/wiki/List%20of%20film...
index                                                            109430
content_embedding     [-0.024242813, 0.019909445, -0.042916078, 0.01...
Name: 0, dtype: object
```

https://arxiv.org/pdf/2212.09741.pdf

Storage

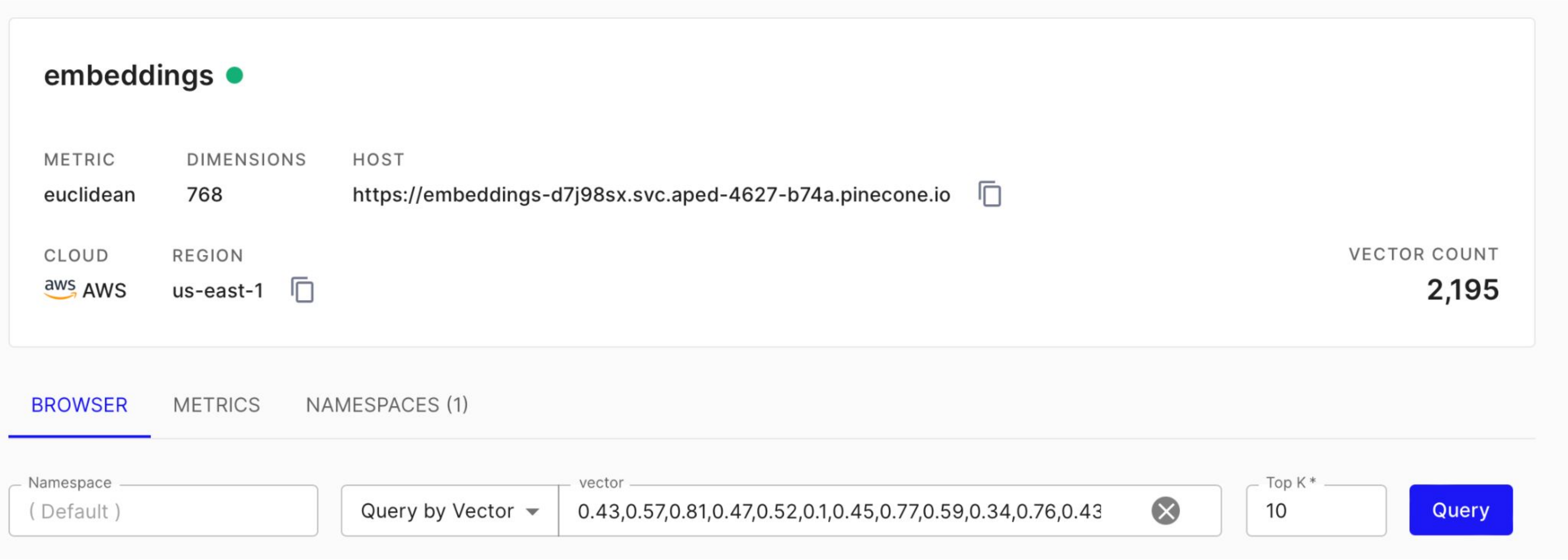
- Vector index (ex. Faiss) and vector database options
- Used Wikipedia dataset



The landscape of vector databases.

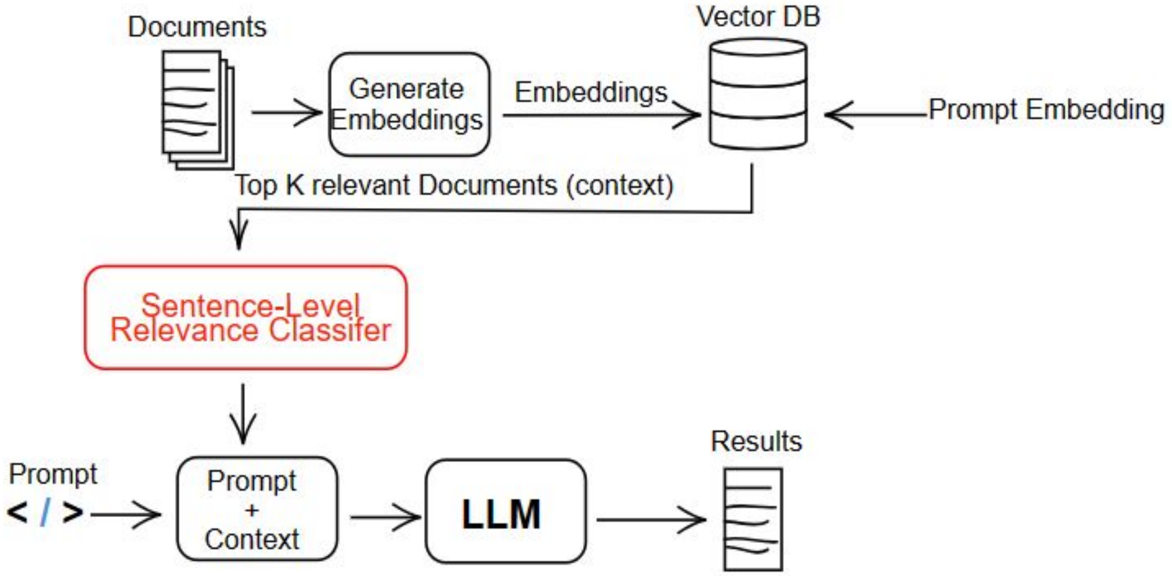
https://blog.det.life/why-you-shouldnt-invest-in-vector-databases-c0cd3f59d23c

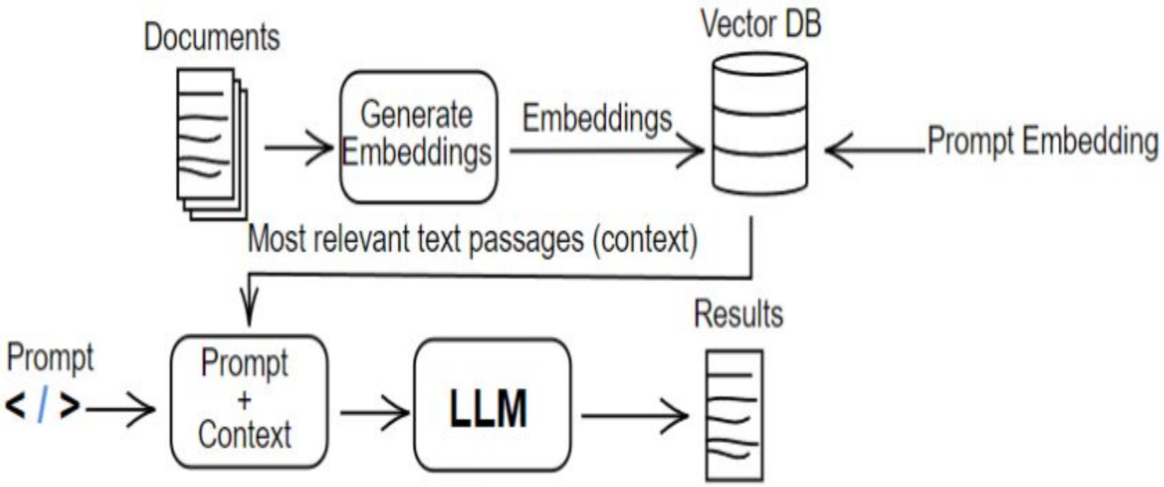
Storage




```
sending upsert requests: 100% 2355/2355 [01:23<00:00, 33.19it/s]
{'upserted_count': 2355}
```

Proposed vs Existing Framework





Wikipedia Document Dataset

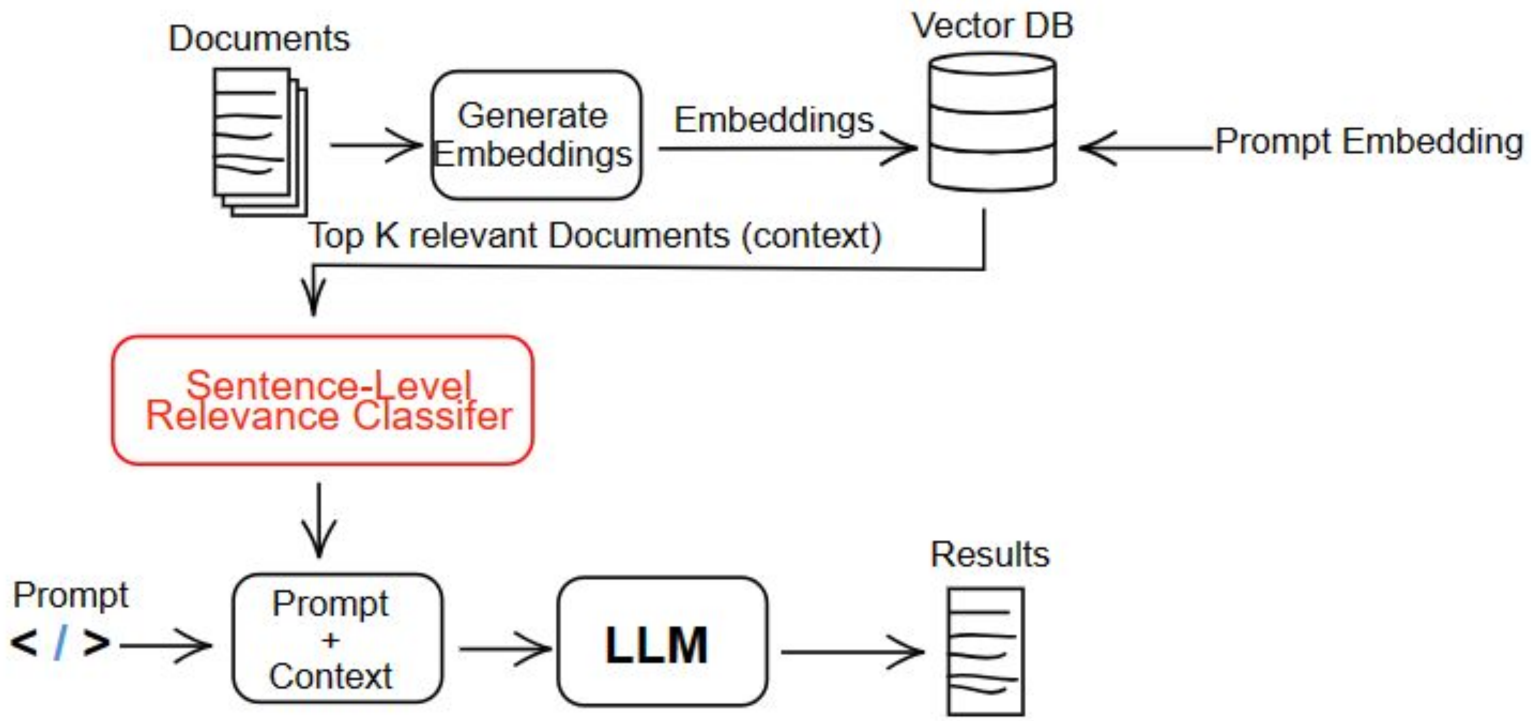
id int32 0 – 1.39M	title string · lengths 1 – 79	text string · lengths 101 – 9.71k	url string · lengths 38 – 44	wiki_id int32 12 – 72M	views float32 1.99k – 5.67k	paragraph_id int32 0 – 603	langs int32 0 – 312	emb sequence
0	Deaths in 2022	The following notable deaths…	https://en.wikipedia.org/wiki?curid=69407798	69,407,798	5,674.449219	0	38	[0.2865696847438812, -0.03181683272123337,…
1	YouTube	YouTube is a global online…	https://en.wikipedia.org/wiki?curid=3524766	3,524,766	5,409.561035	0	184	[-0.0968938171863555, 0.1619211882352829,…
2	YouTube	In October 2006, YouTube was bough…	https://en.wikipedia.org/wiki?curid=3524766	3,524,766	5,409.561035	1	184	[0.1302049309015274, 0.265736848115921,…
3	YouTube	Since its purchase by Google, YouTub…	https://en.wikipedia.org/wiki?curid=3524766	3,524,766	5,409.561035	2	184	[-0.0979125723242759, 0.13586106896400452,…
4	YouTube	YouTube has had an unprecedented…	https://en.wikipedia.org/wiki?curid=3524766	3,524,766	5,409.561035	3	184	[-0.2641527056694031, 0.06968216598033905,…

Unnamed: 0	query	long_answer	short_answer	title	bert_title	abstract	content	url	index	content_embedding
24	what are the toll roads called in mexico	This is a list of autopistas , or tolled (quota) highways , in Mexico . T	autopistas	List of Mexican aut	list of mexican autopistas	This is a list c	Many federa	https://en.wi	109631	5.41652627e-02 6.37640506e-02 -5.65397879e-03 4.89284545e-02
25	what are the top five wine producing states	2016 production of still wine State Production (gal) Production	California|Washington|New York|Penr	American wine	american wine	American wir	The North Ar	https://en.wi	98864	[[-7.31019536e-03 -1.26333470e-02 -1.06264362e-02 1.31341349e-02
26	who sings the theme song for living single	Living Single Season 1 DVD cover Created by Yvette Lee Bowser	performed by	Living Single	living single	Living Single	Throughout	https://en.wi	15276	[[-0.01875372 -0.00664845 -0.01411123 0.01214166 0.03976656 0.0153
27	what type of reproduction do whiptail lizards use	summer , and hatching approximately eight weeks later . The New M	-	New Mexico whipta	new mexico whiptail	Cnemidophc	The New Me:	https://en.wi	103123	[[-2.50664949e-02 -7.14592077e-03 -2.15893276e-02 3.62700666e-03
28	where was the summer olympics held in 2012	The 2012 Summer Olympics , formally the Games of the XXX Olympi	Olympiad	2012 Summer Olyn	2012 summer olympics	The 2012 Su	Following a bid headed by former Olympic champion Sebastian Coe and then , Mayor of London Ken Livingstone , Londo			

https://huggingface.co/datasets/Cohere/wikipedia-22-12-en-embeddings/viewer/default/train

Methodology

- Framework:



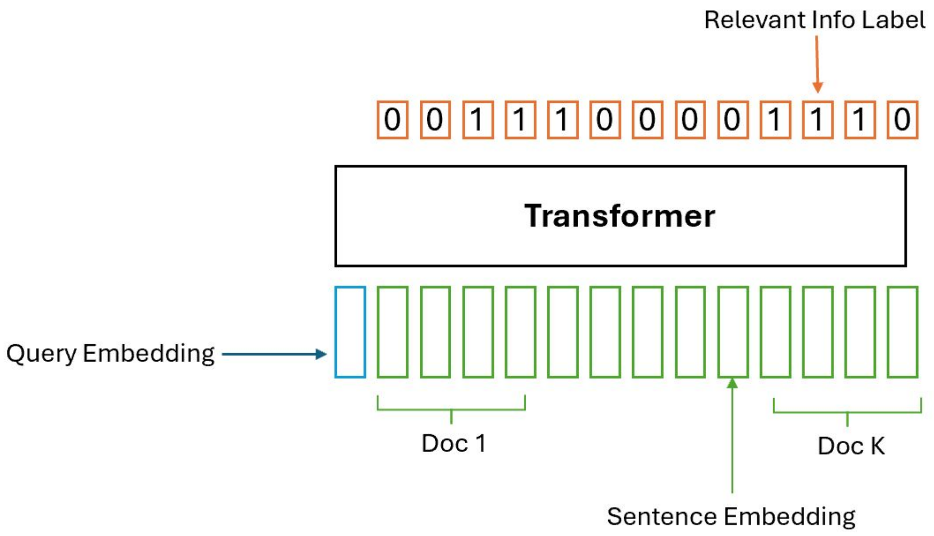
Methodology

- Motivation:
  - RAG relies on highly relevant information.
  - Traditional methods like reranking and truncations have some limitations

- We propose a model to identify most relevant information.

Proposed Method

- SLRC: Sentence-Level Relevance Classifier
- Input:
  - Query Embedding and sentences embedding of Top K documents
- Output:
  - Relevant info label: if the sentences is relevant with the query.



Evaluation

Natural Questions Benchmark

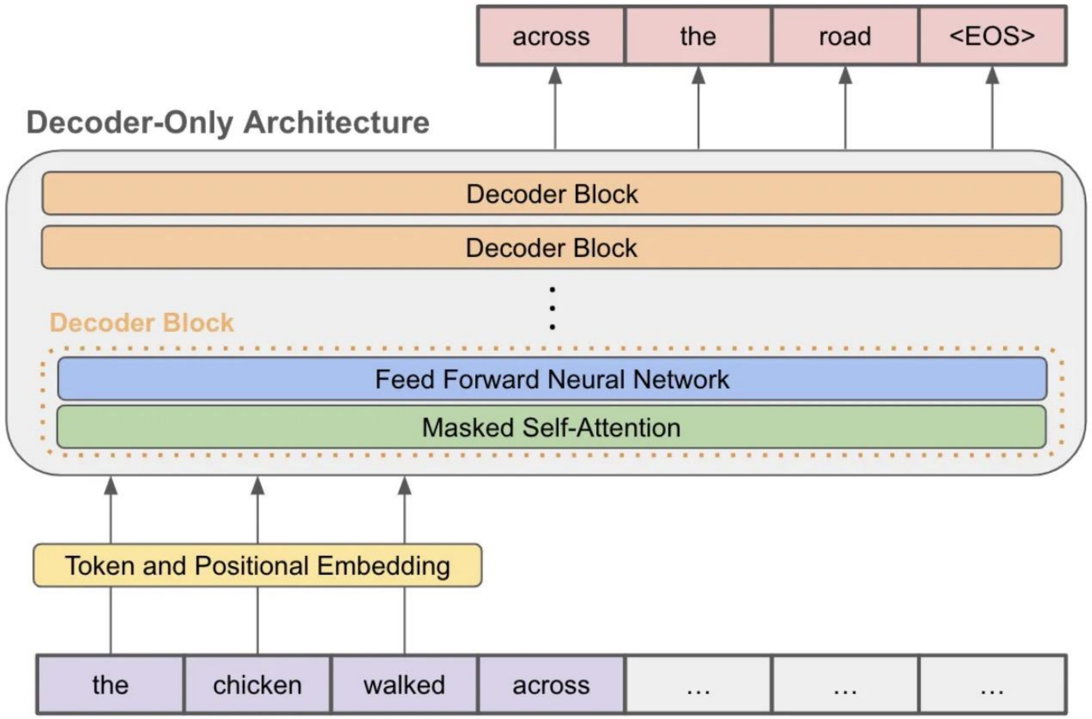
Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".

Short answer: jet-black

Example 1
Question: who played will on as the world turns **Long answer:** William "Will" Harold Ryan Munson is a fictional character on the CBS soap opera As the World Turns. He was portrayed by Jesse Soffer on recurring basis from September 2004 to March 2005, after which he got a contract as a regular. Soffer left the show on April 4, 2008 and made a brief return in July 2010. **Judgment:** Correct. **Justification:** It is clear beyond a reasonable doubt that the answer is correct.

https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/1f7b46b5378d757553d3e92ead36bda2e4254244.pdf

LLM (gpt-2)



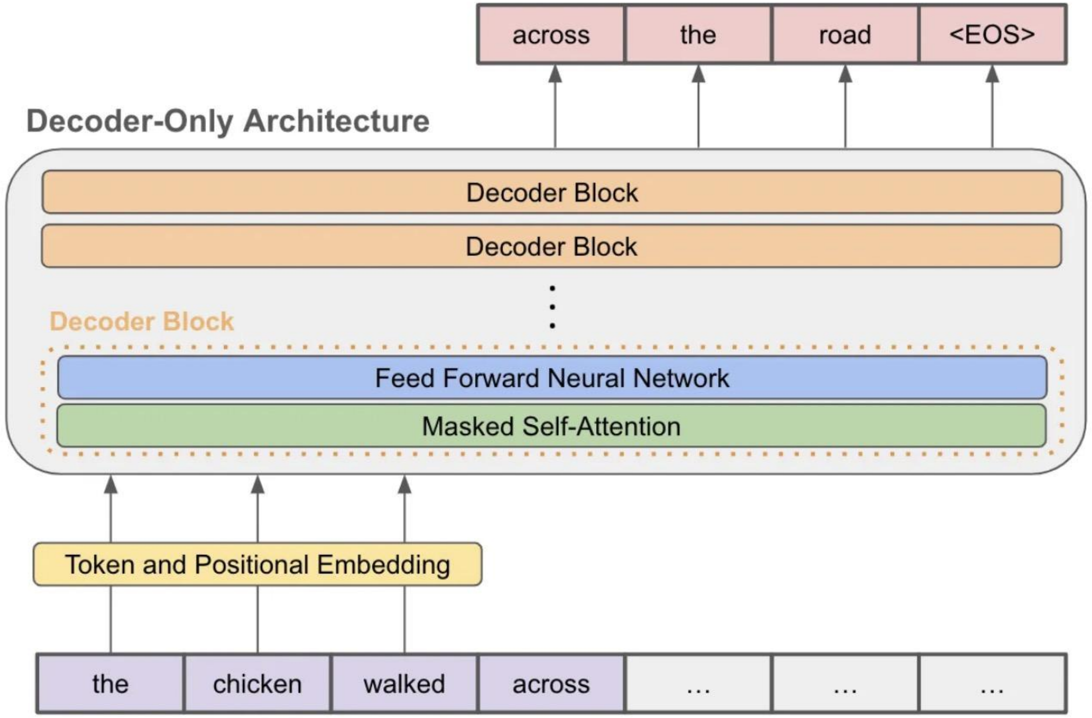
Decoder-only transformer architecture (created by author)

GPT-2 is a decoder-only transformer architecture. It removes the following components from the transformer:

- The entire encoder module
- All encoder-decoder self-attention modules in the decoder

https://towardsdatascience.com/language-models-gpt-and-gpt-2-8bdb9867c50a

LLM (gpt-2)



Decoder-only transformer architecture (created by author)

https://towardsdatascience.com/language-models-gpt-and-gpt-2-8bdb9867c50a

After these components have been removed, each layer of the decoder simply consists of

- a masked self-attention layer followed by a feed forward neural network. Stacking several of such layers on top of each other forms a deep, decoder-only transformer architecture, such as those used for GPT or GPT-2.

Evaluation Metrics

We choose two metrics to measure the similarity of the machine-translated text to a set of high quality reference translations. Both metrics range from 0 to 1.

- BLEU_SCORE: A precision based measure.
- ROUGE_SCORE: The harmonic mean of recall and precision.
- Perplexity score ?

Evaluation Benchmarks

We choose the RAG and GPT-2 as the benchmarks where GPT-2 is pre-trained on large scale datasets.

We compare the performance of:

1. Pretrained GPT-2 model
2. Pretrained GPT-2 model + RAG prompts (top-3 documents)
3. Fine-tuned GPT-2 model using RAG prompts (top-3 documents)
4. LLaMa 7B + RAG prompts
5. LLaMa 7B

Evaluation settings

We customize the dataset by selecting 2335 query&answer pairs from NQDataset. The dataset is splitted into two parts:

- Train Split: 1335 query&answer pairs
- Test Split: 1000 query&answer pairs

What's more, the indices of top-3 document is added for each query&answer pairs.

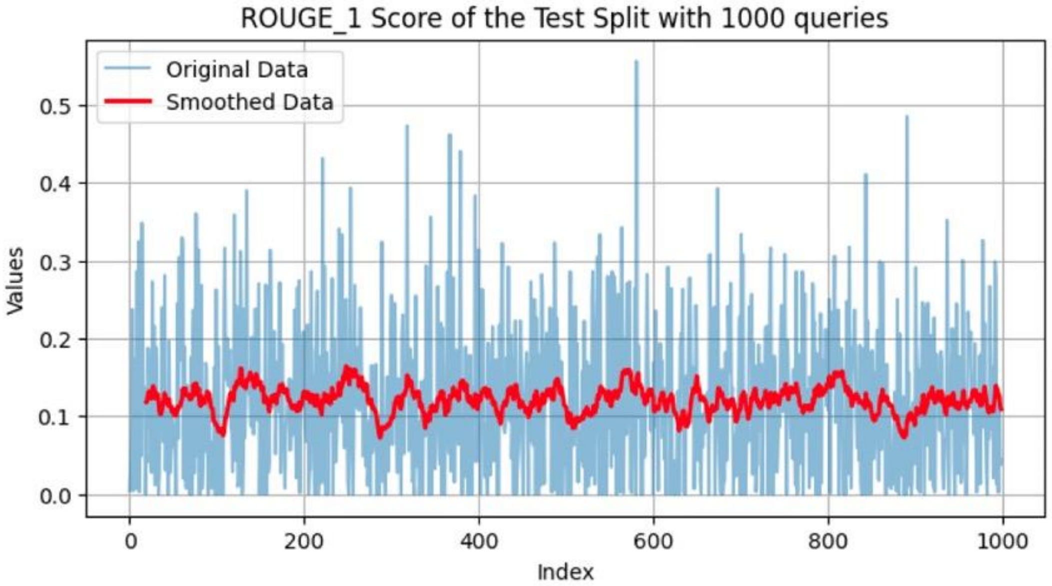
query	long_answer	Topk_indices
wolf of wall street number of f words	Film Year Fuck count Minutes Uses / minute Source	[0 178 684]

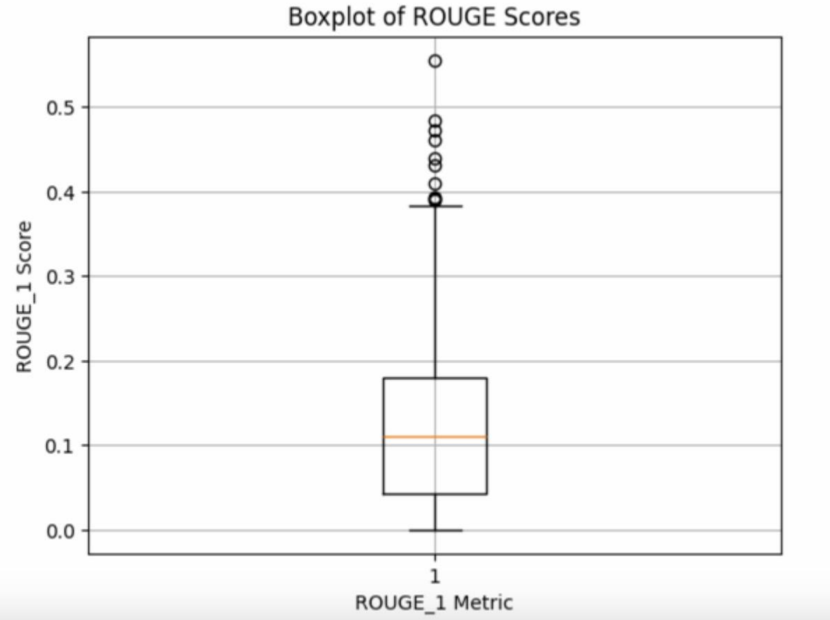
Example of our customized dataset

Evaluation Results — ROUGE of EXP1

The following results show the ROUGE 1 score and BLEU score of the pre-trained GPT-2 model on the test split of our customized dataset.

It shows that the overall performance of pre-trained GPT-2 is bad.

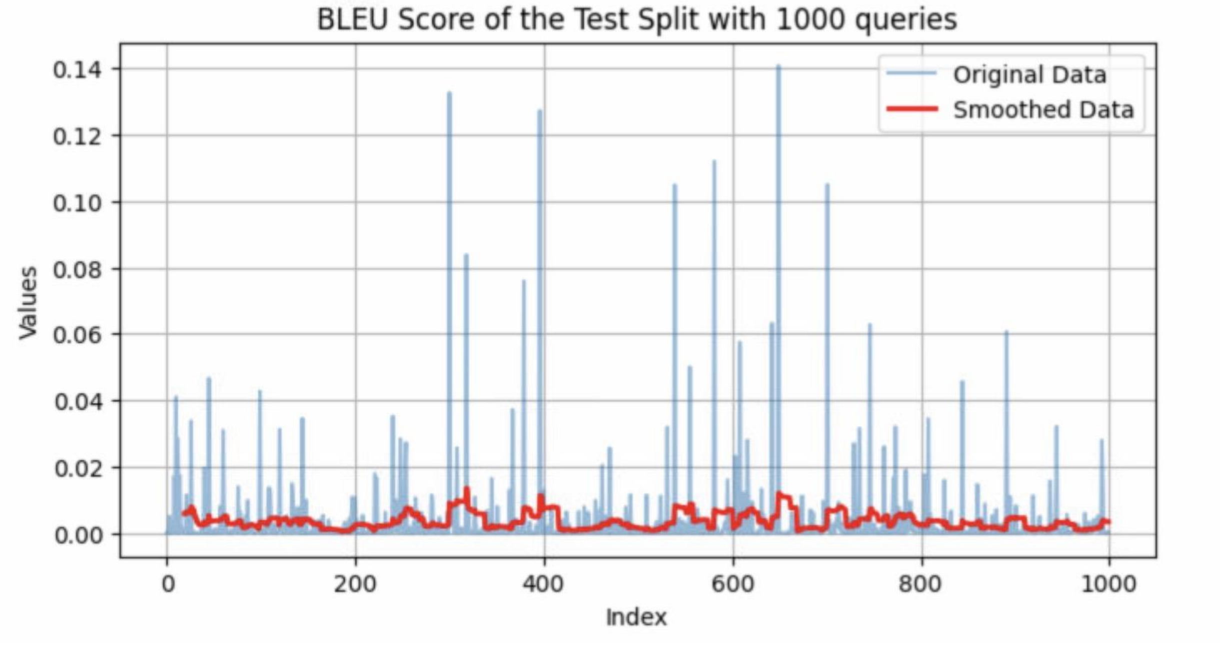


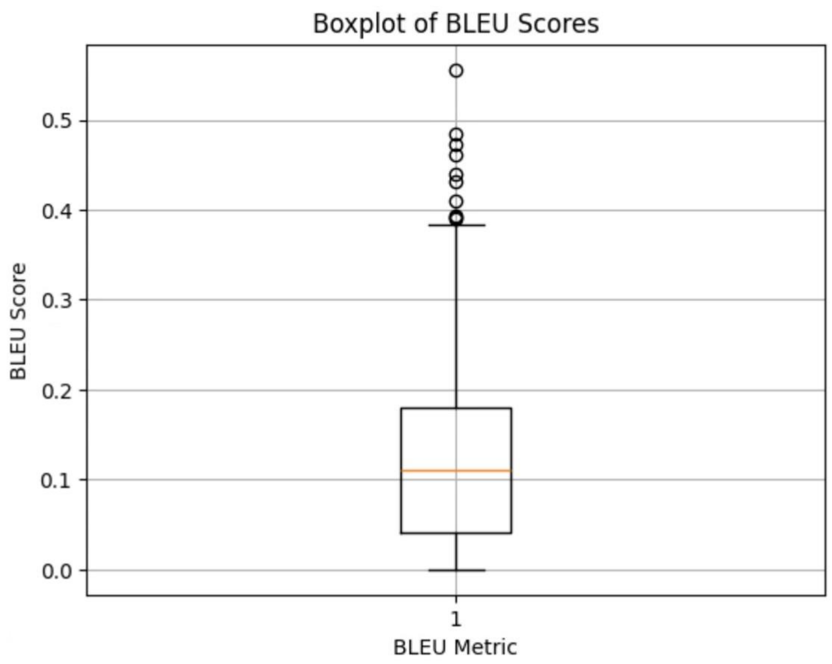


Evaluation Results — BLEU of EXP1

The following results show the ROUGE 1 score and BLEU score of the pre-trained GPT-2 model on the test split of our customized dataset.

It shows that the overall performance of pre-trained GPT-2 is bad.

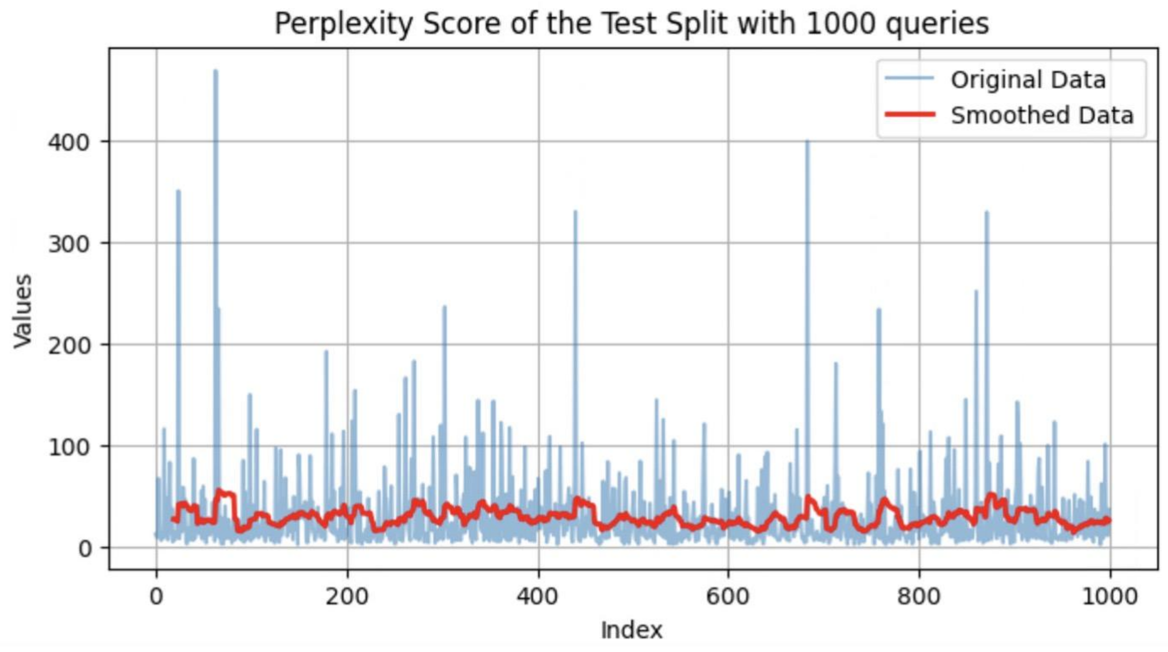


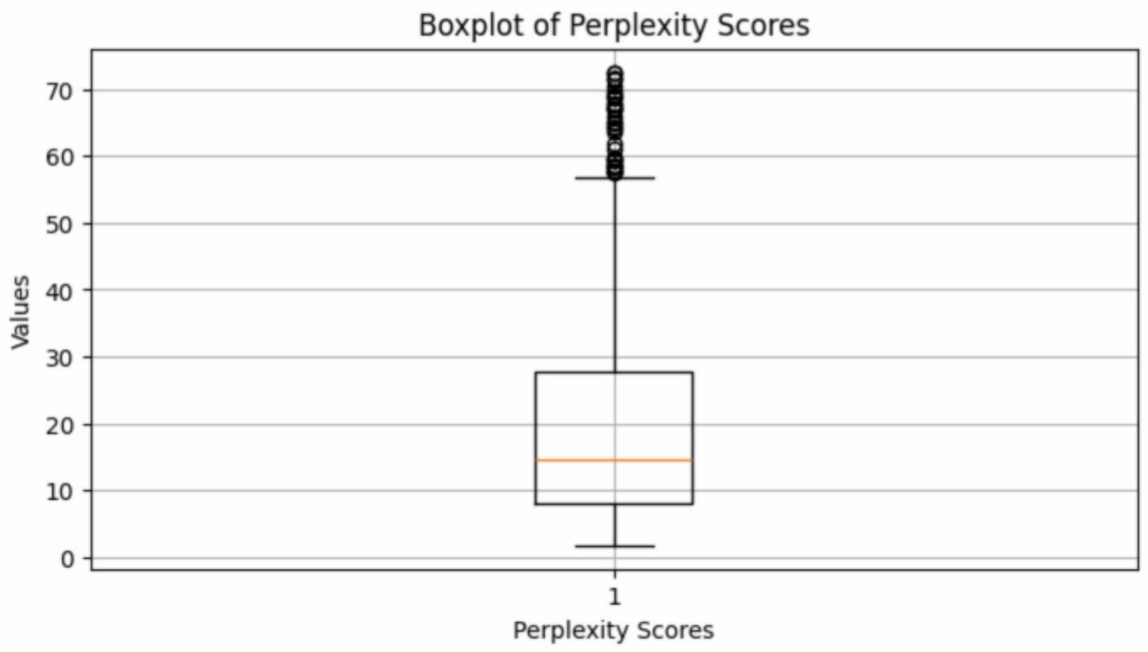


Evaluation Results — Perplexity of EXP1

The following results show the ROUGE 1 score and BLEU score of the pre-trained GPT-2 model on the test split of our customized dataset.

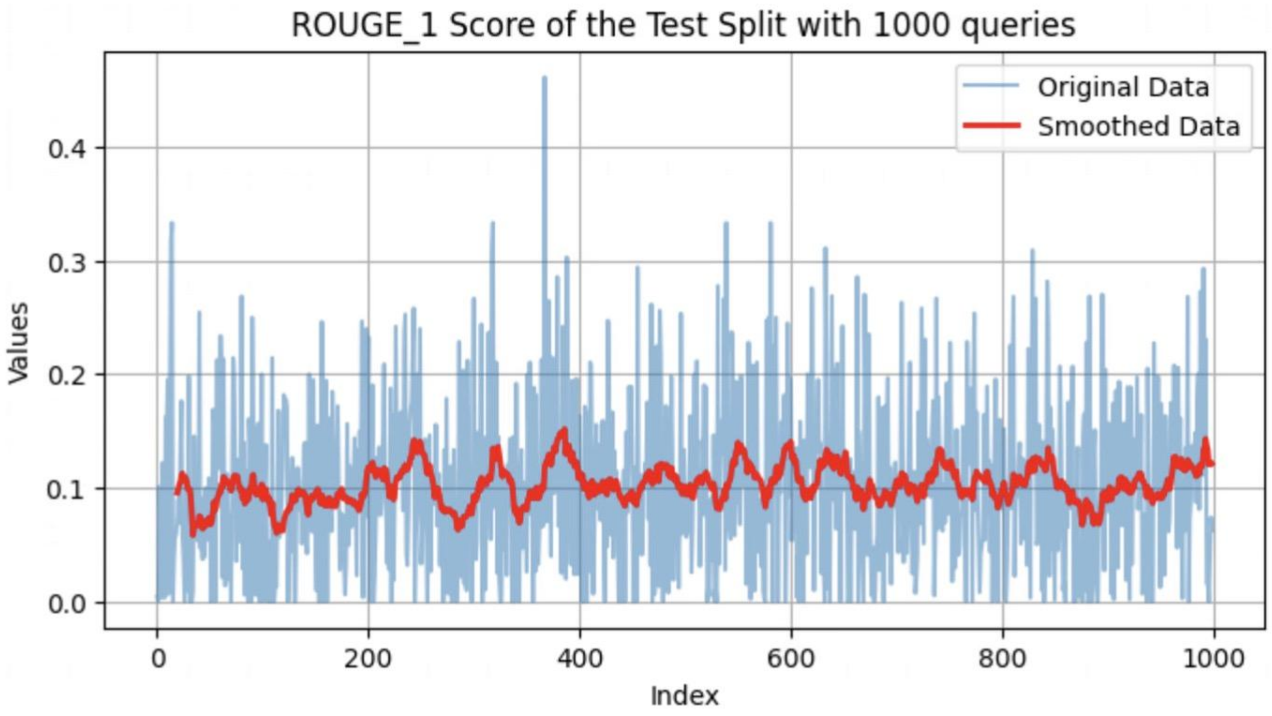
It shows that the overall performance of pre-trained GPT-2 is bad.

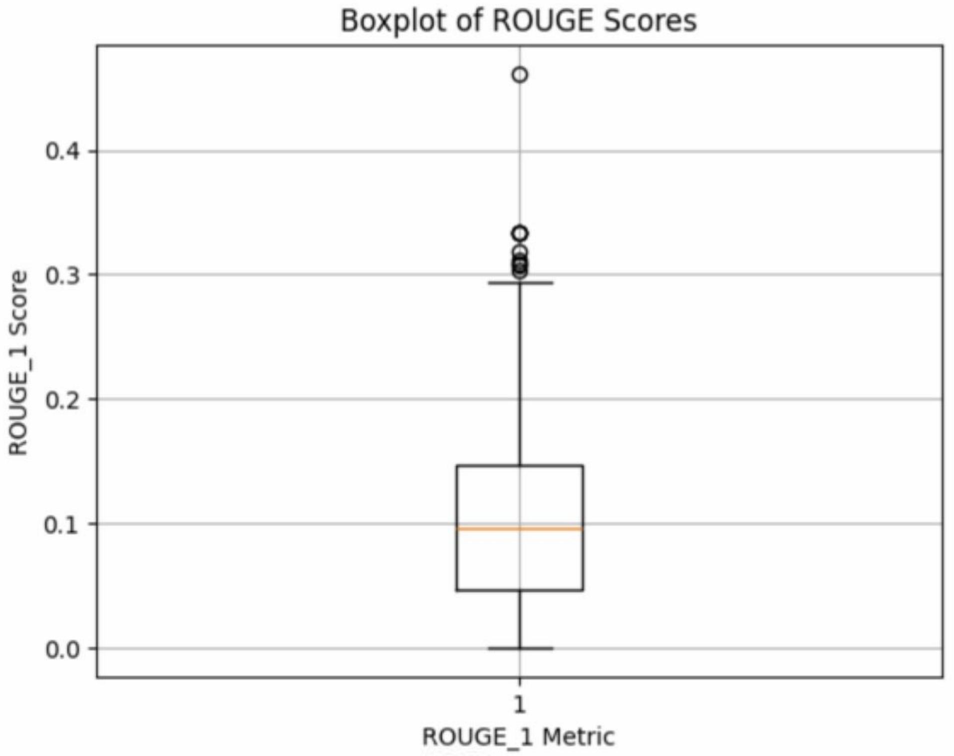




Evaluation Results — ROUGE of EXP2

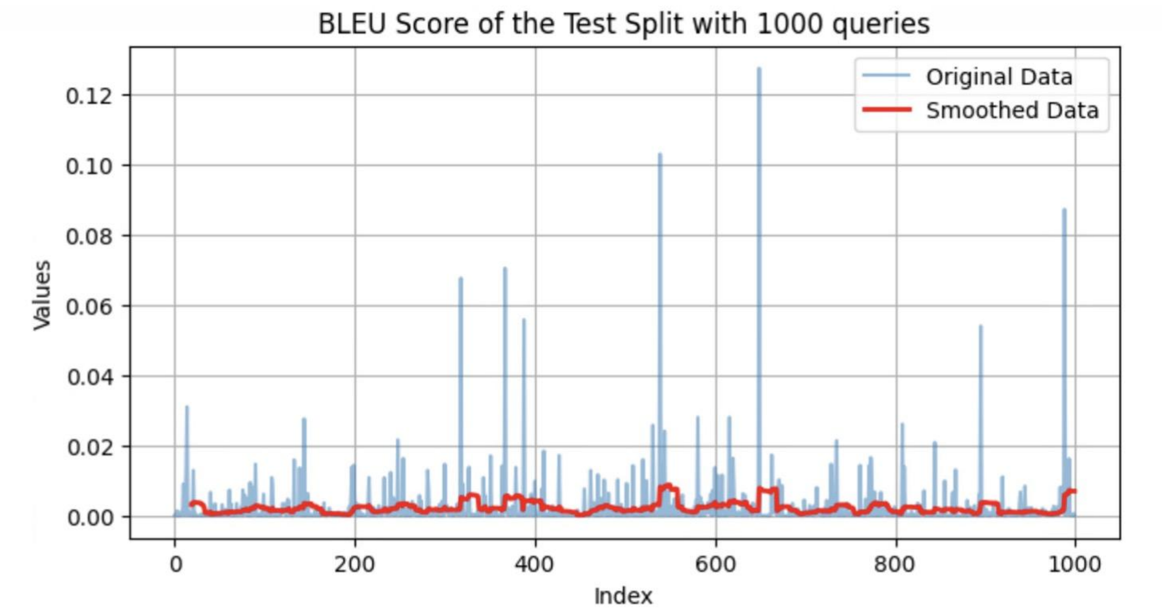
In terms of the bad performance of pretrained GPT-2 mode, we use RAG prompts (top-3 documents) to help improve the scores.

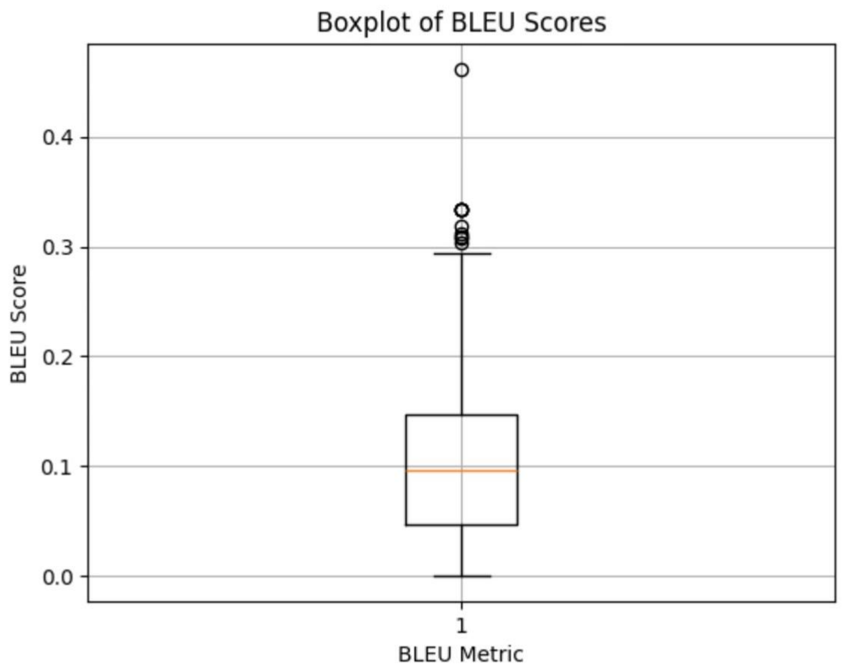




Evaluation Results — BLEU of EXP2

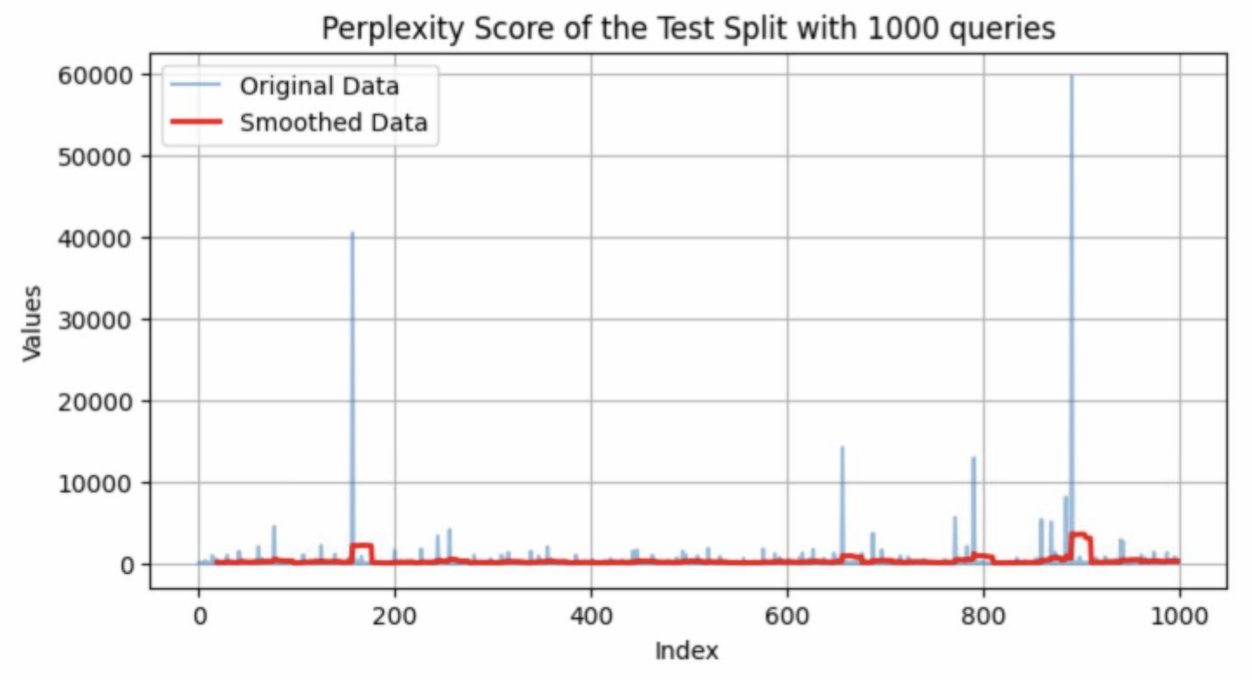
In terms of the bad performance of pretrained GPT-2 mode, we use RAG prompts (top-3 documents) to help improve the scores.

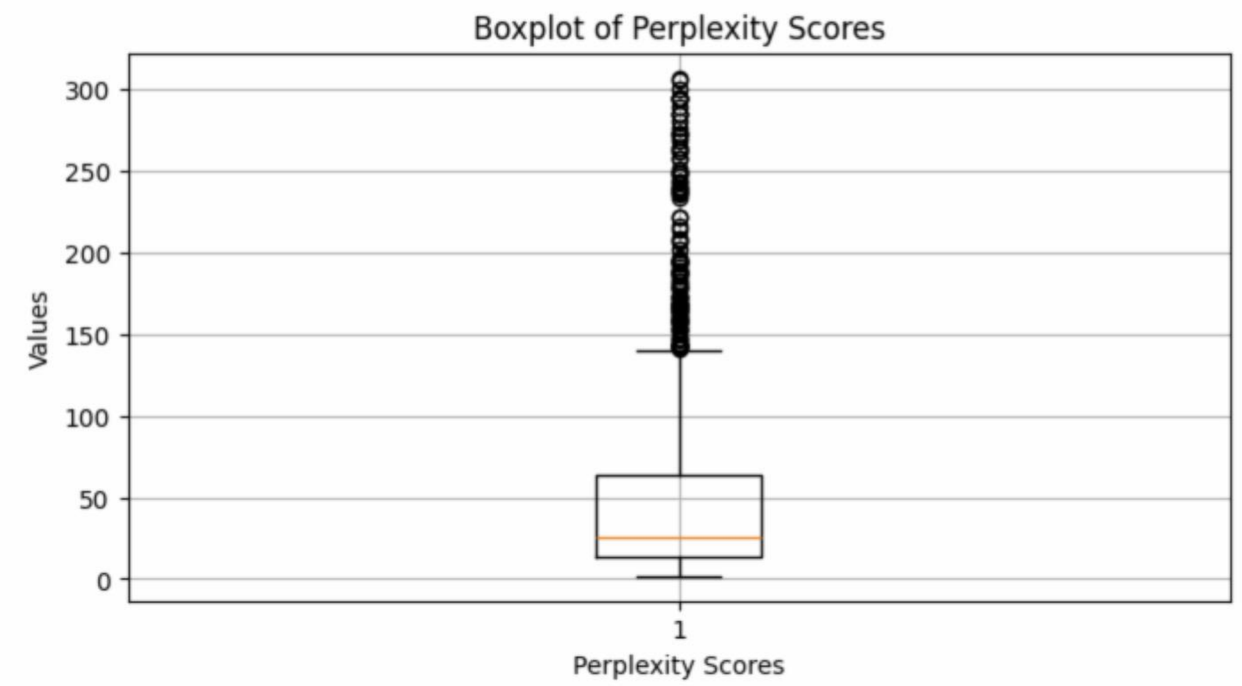




Evaluation Results — Perplexity of EXP2

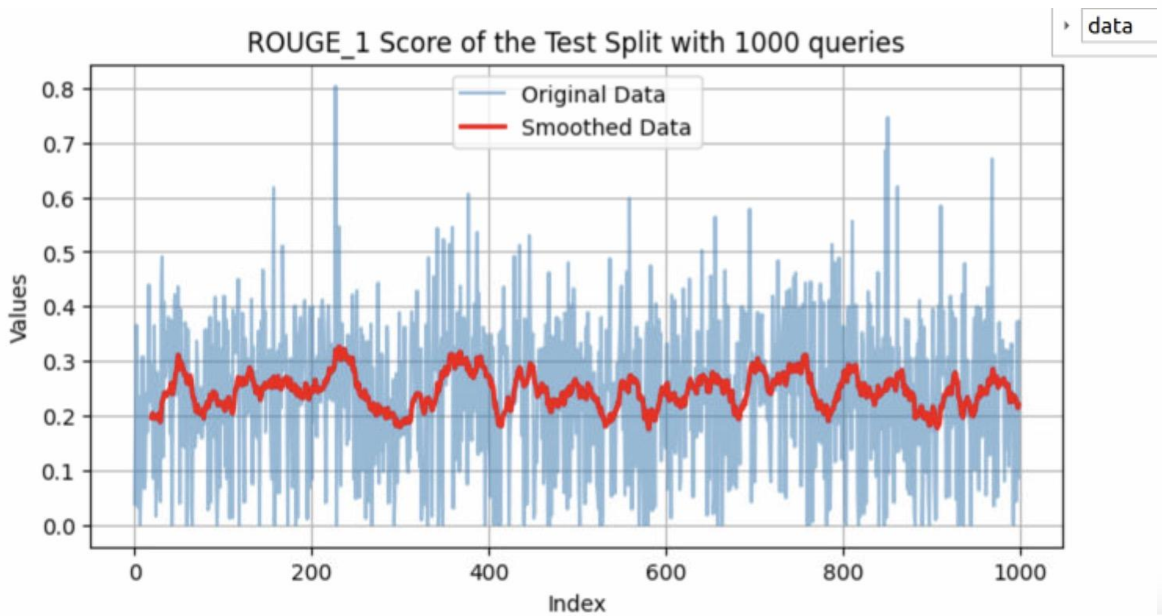
In terms of the bad performance of pretrained GPT-2 mode, we use RAG prompts (top-3 documents) to help improve the scores.

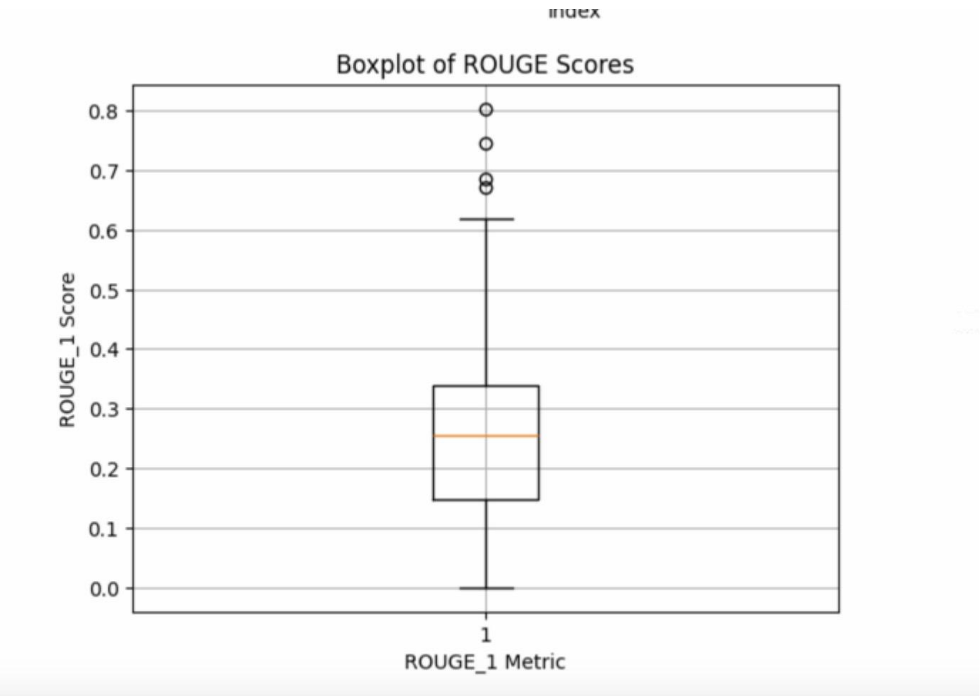




Evaluation Results — ROUGE of EXP3

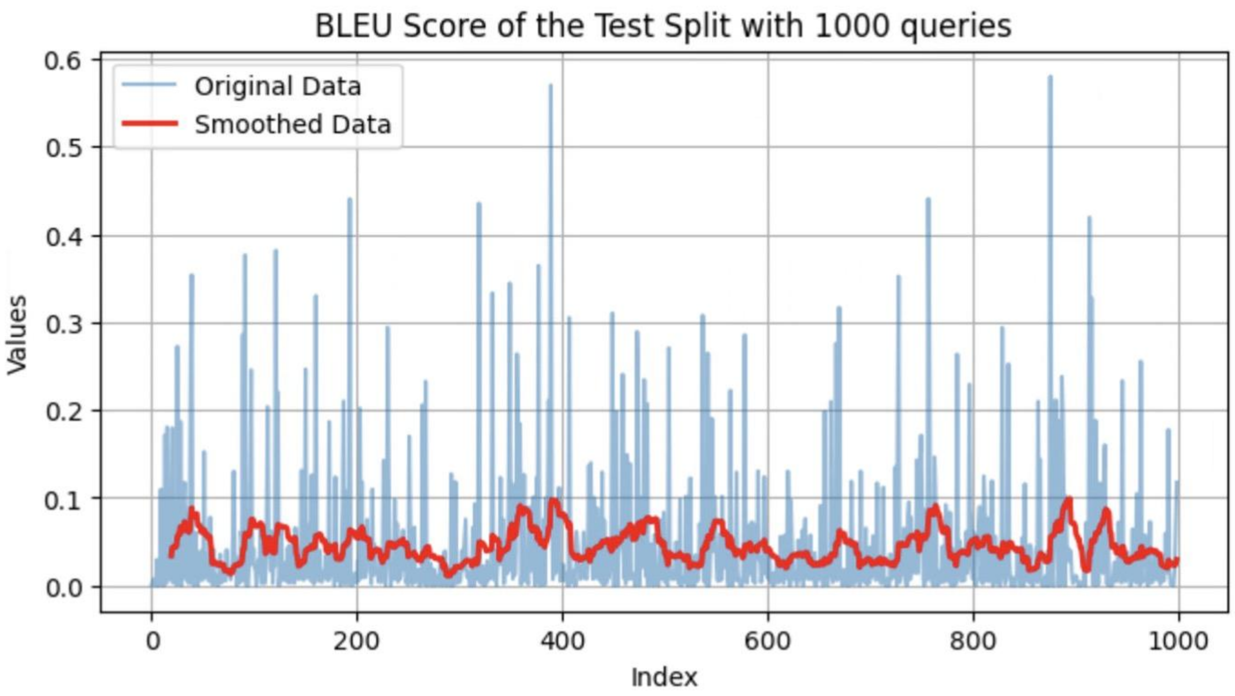
In terms of the bad performance of pretrained GPT-2 mode, we use a classifier to select the most relevant sentences from top-3 documents as the prompt to help improve the scores.

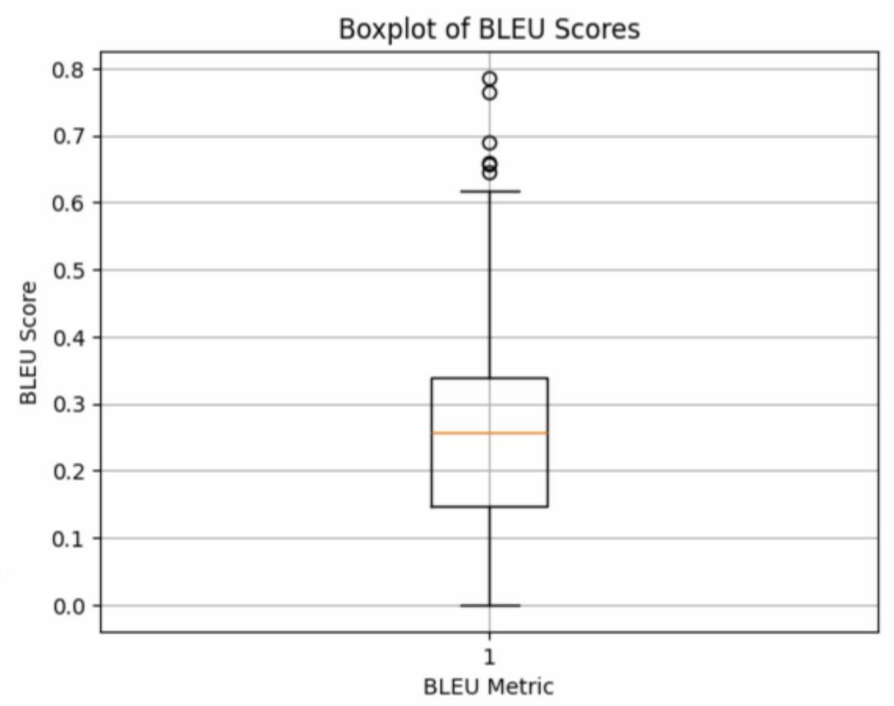




Evaluation Results — BLEU of EXP3

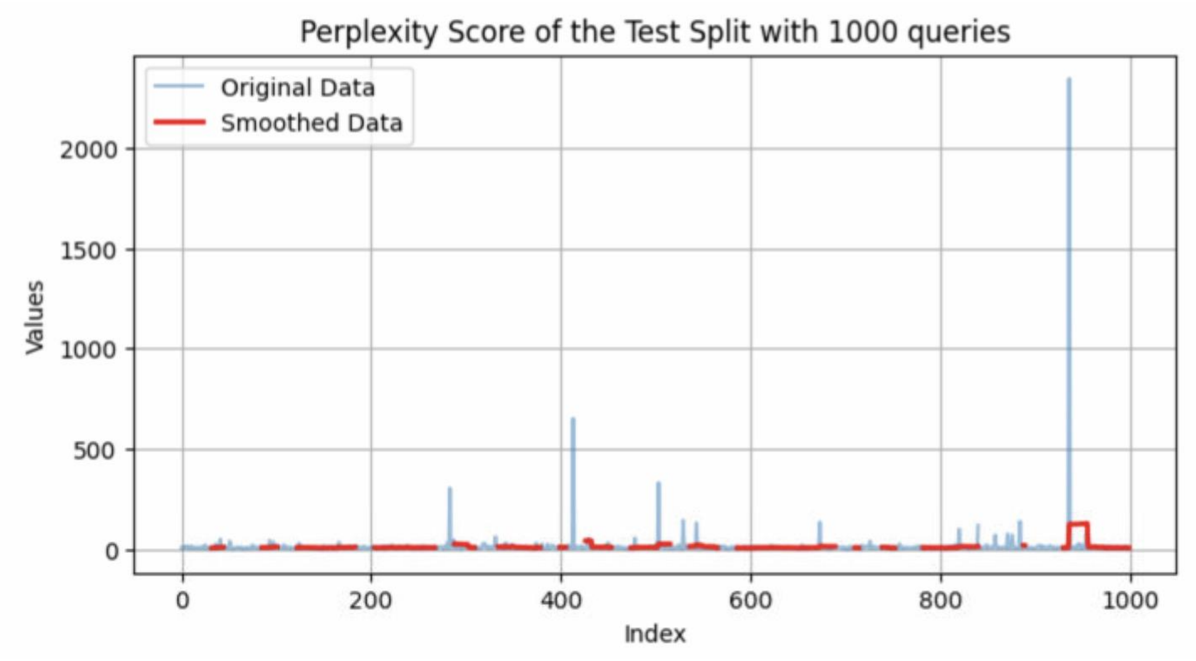
In terms of the bad performance of pretrained GPT-2 mode, we use a classifier to select the most relevant sentences from top-3 documents as the prompt to help improve the scores.

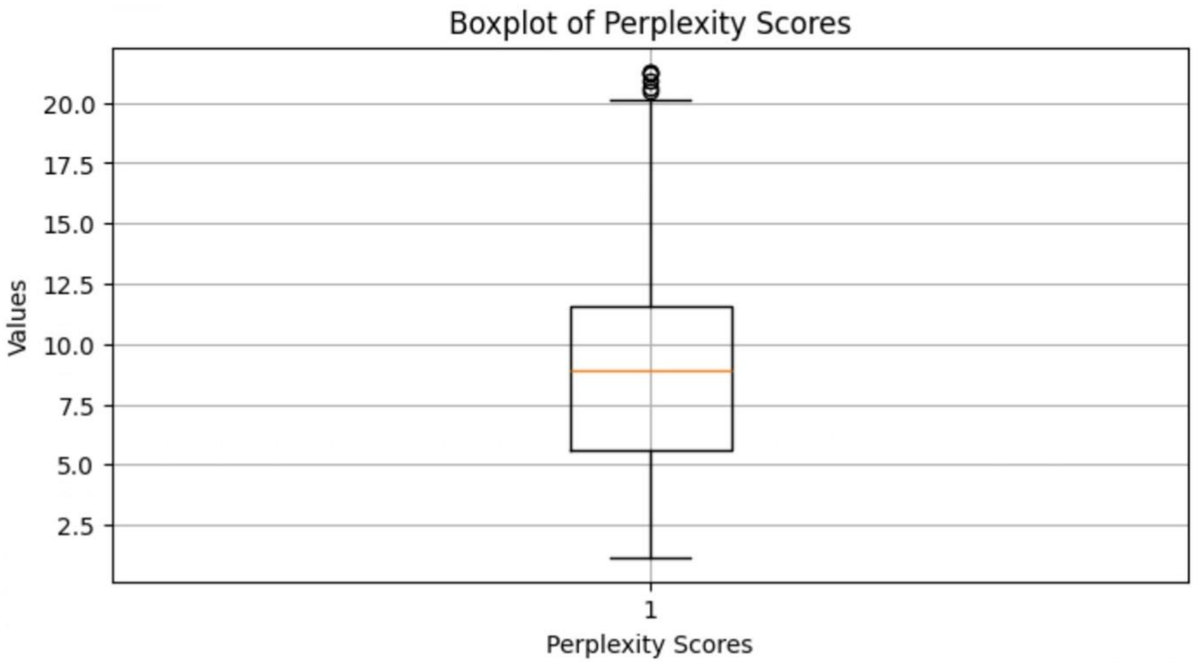




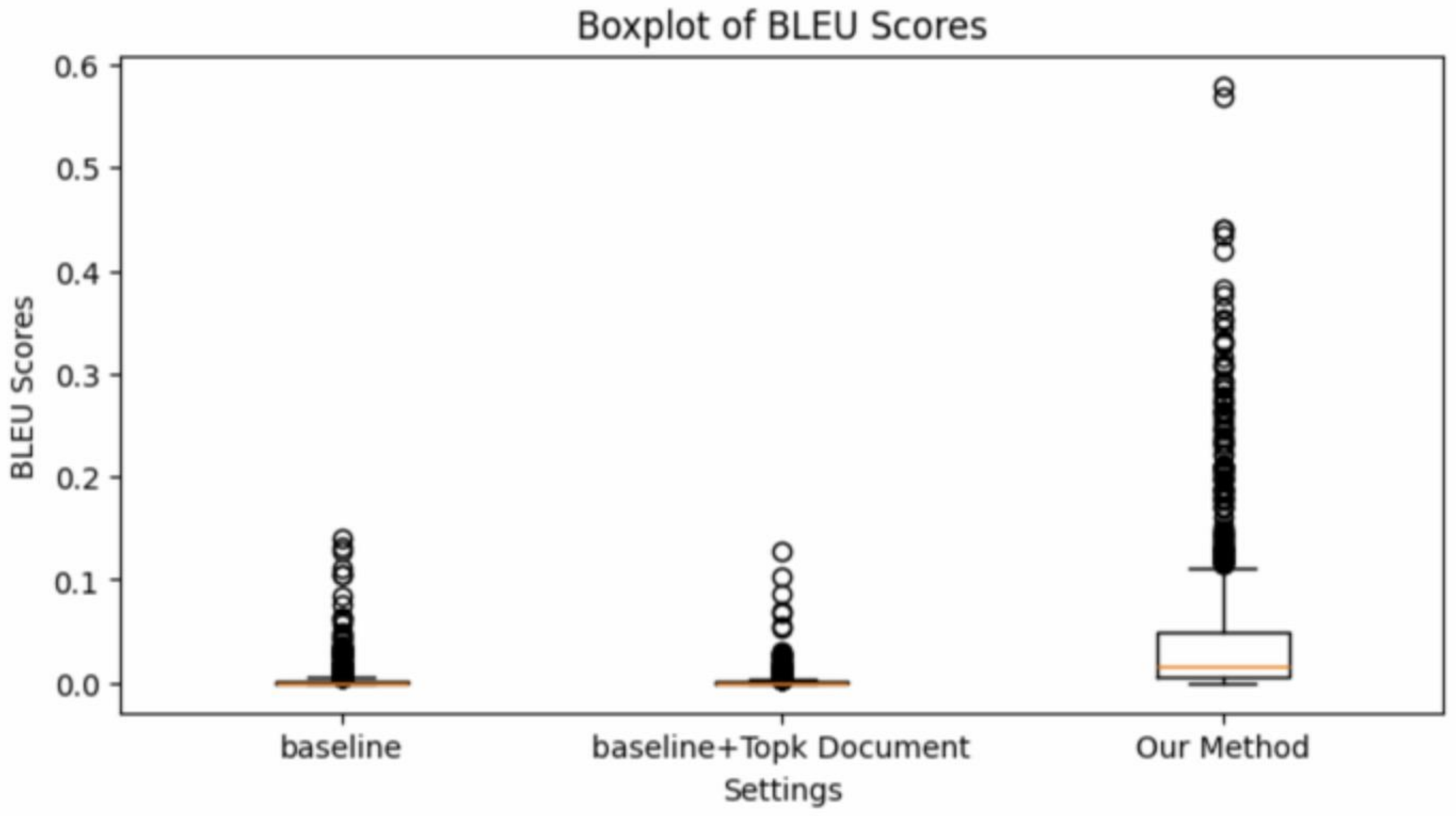
Evaluation Results — Perplexity of EXP3

In terms of the bad performance of pretrained GPT-2 mode, we use a classifier to select the most relevant sentences from top-3 documents as the prompt to help improve the scores.

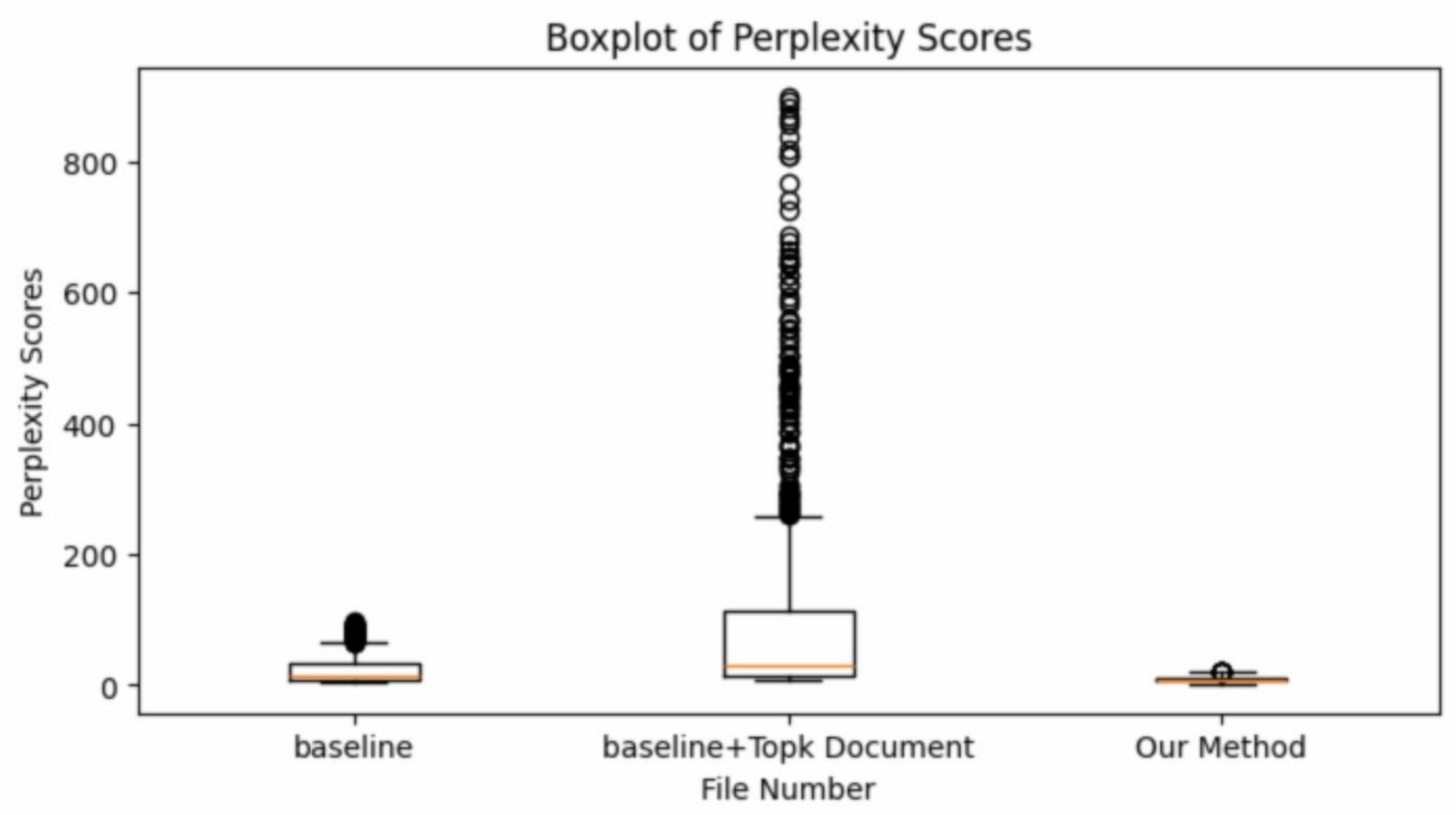




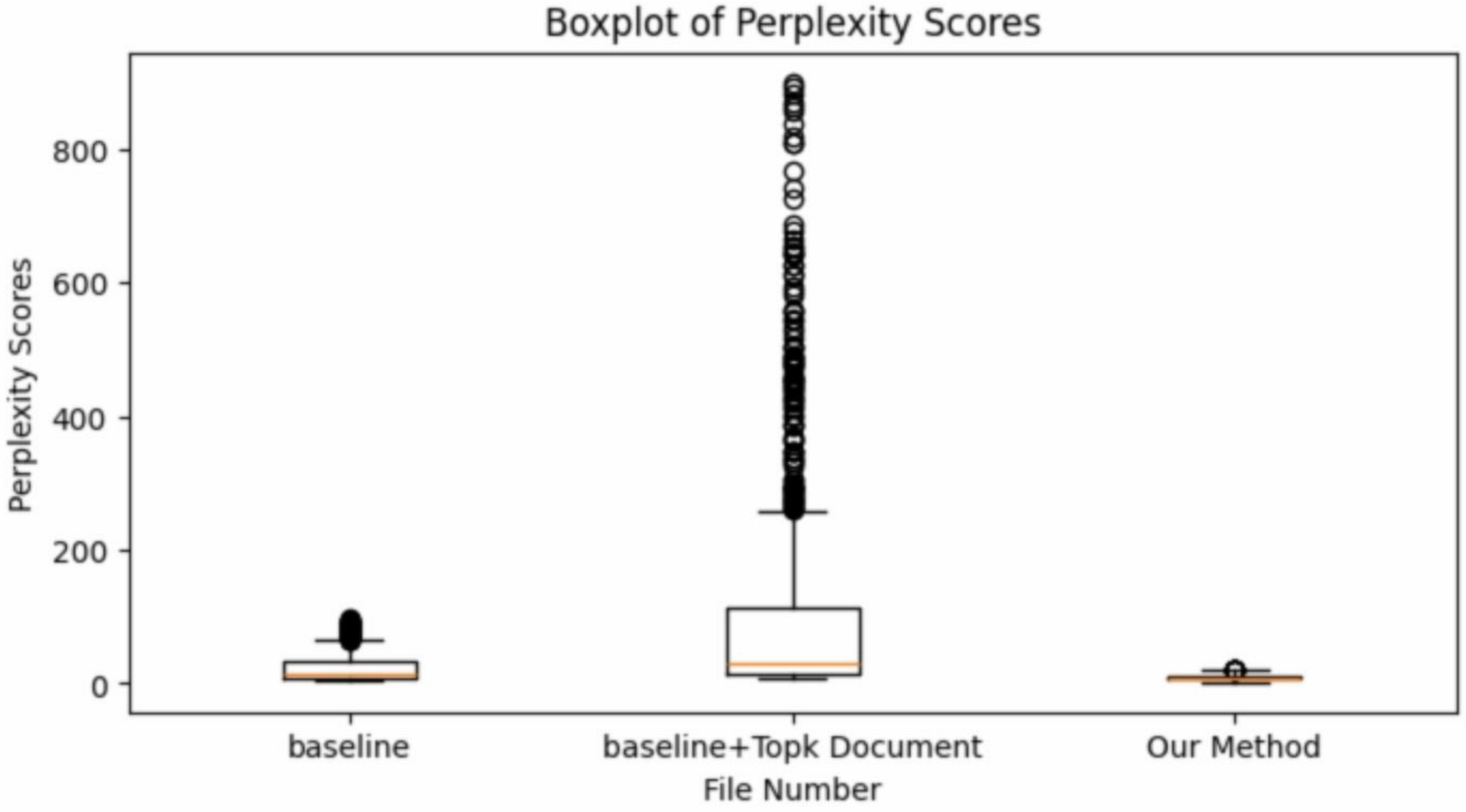
Evaluation Results — BLEU Comparison of GPT2



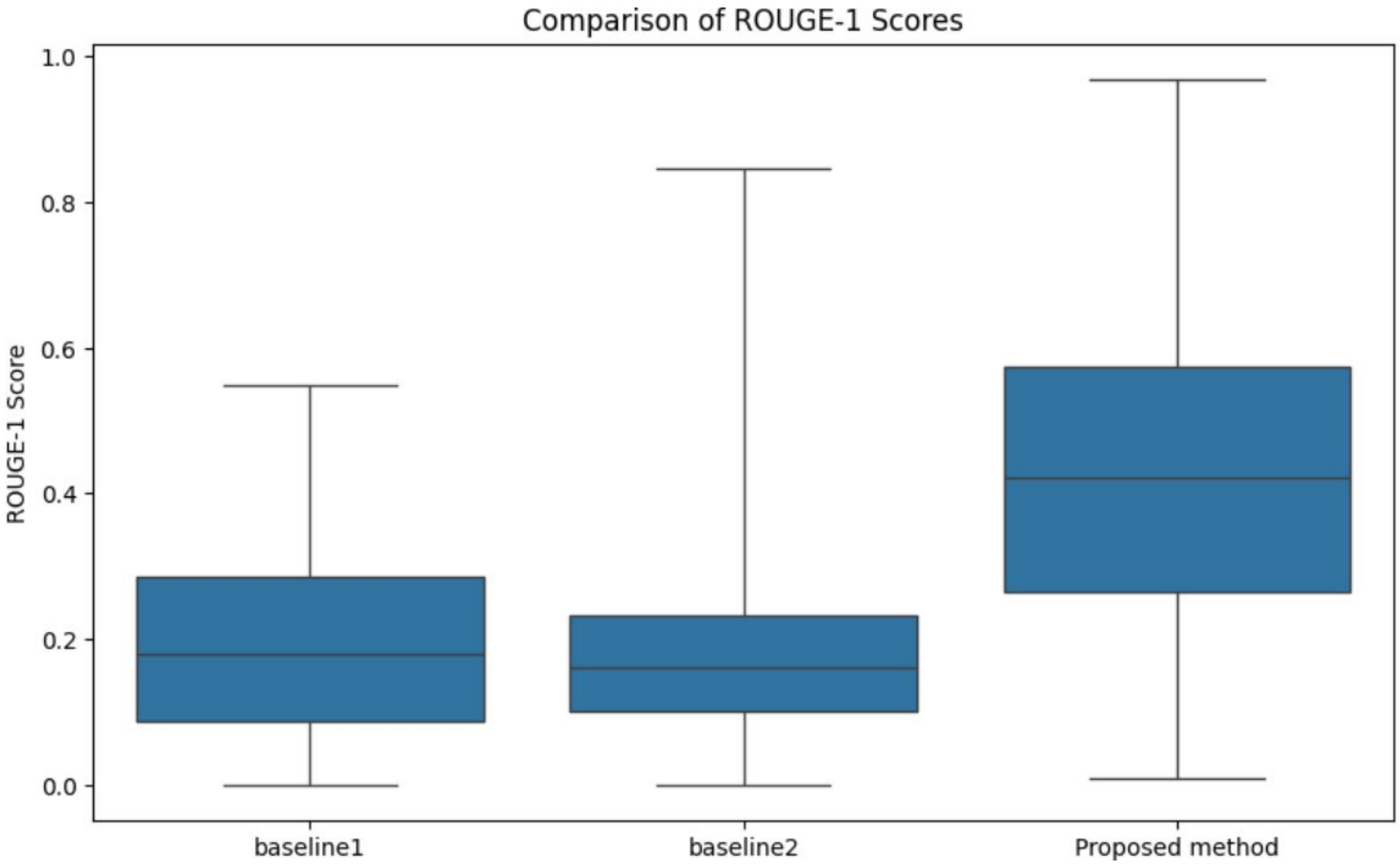
Evaluation Results — ROUGE Comparison of GPT2



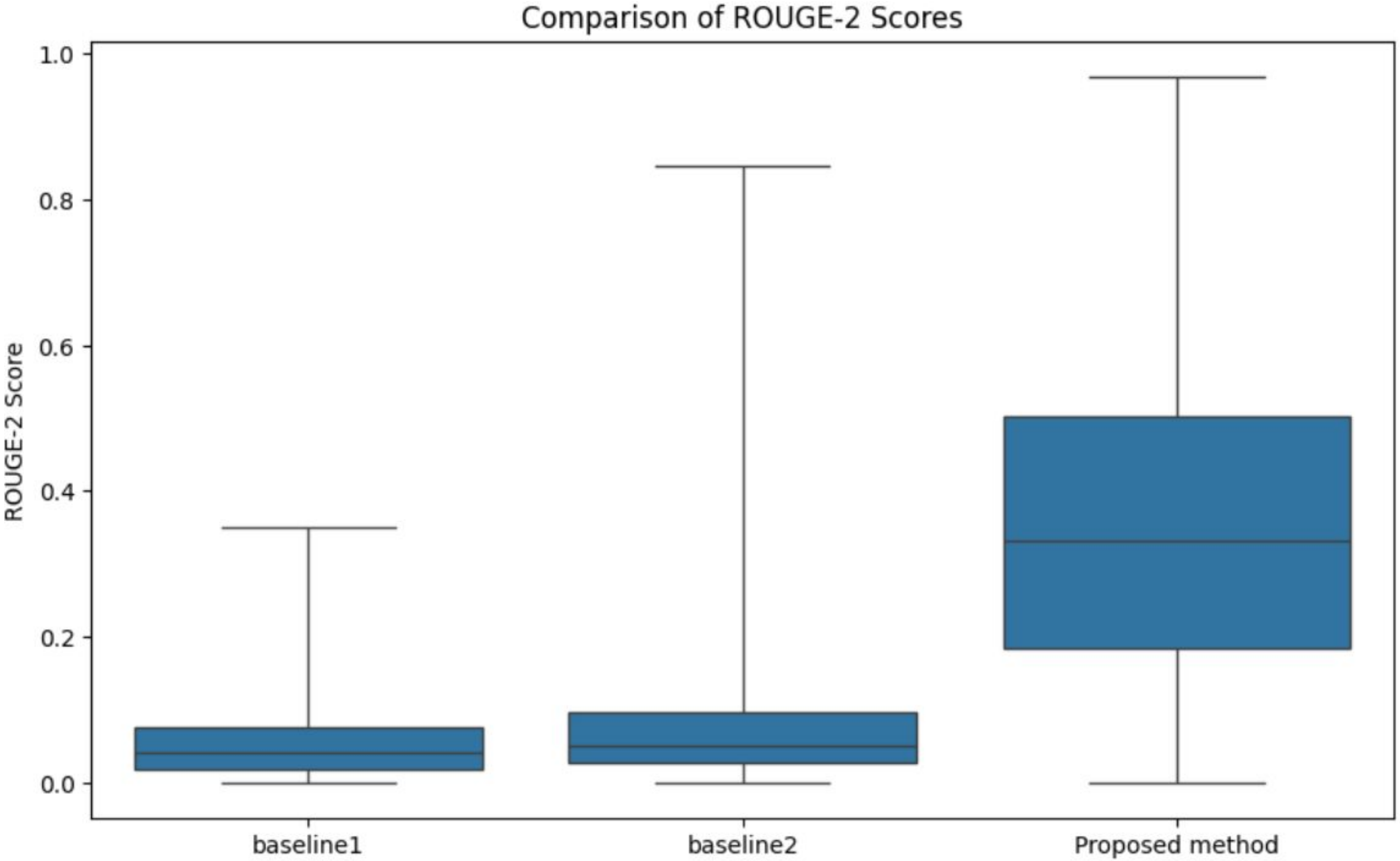
Evaluation Results — Perplexity Comparison of GPT2



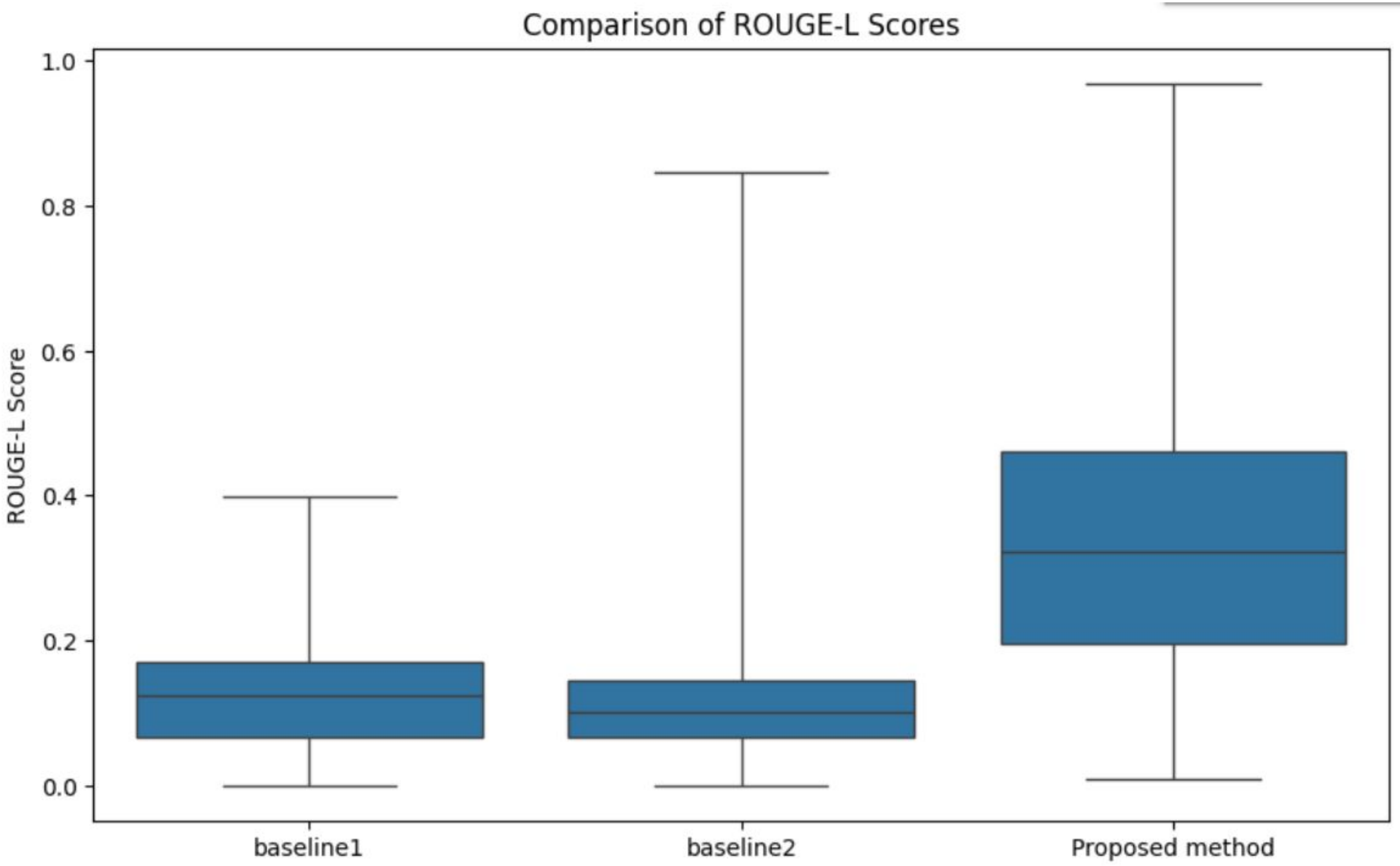
Evaluation Results - Rouge1 Comparison on LLaMa2



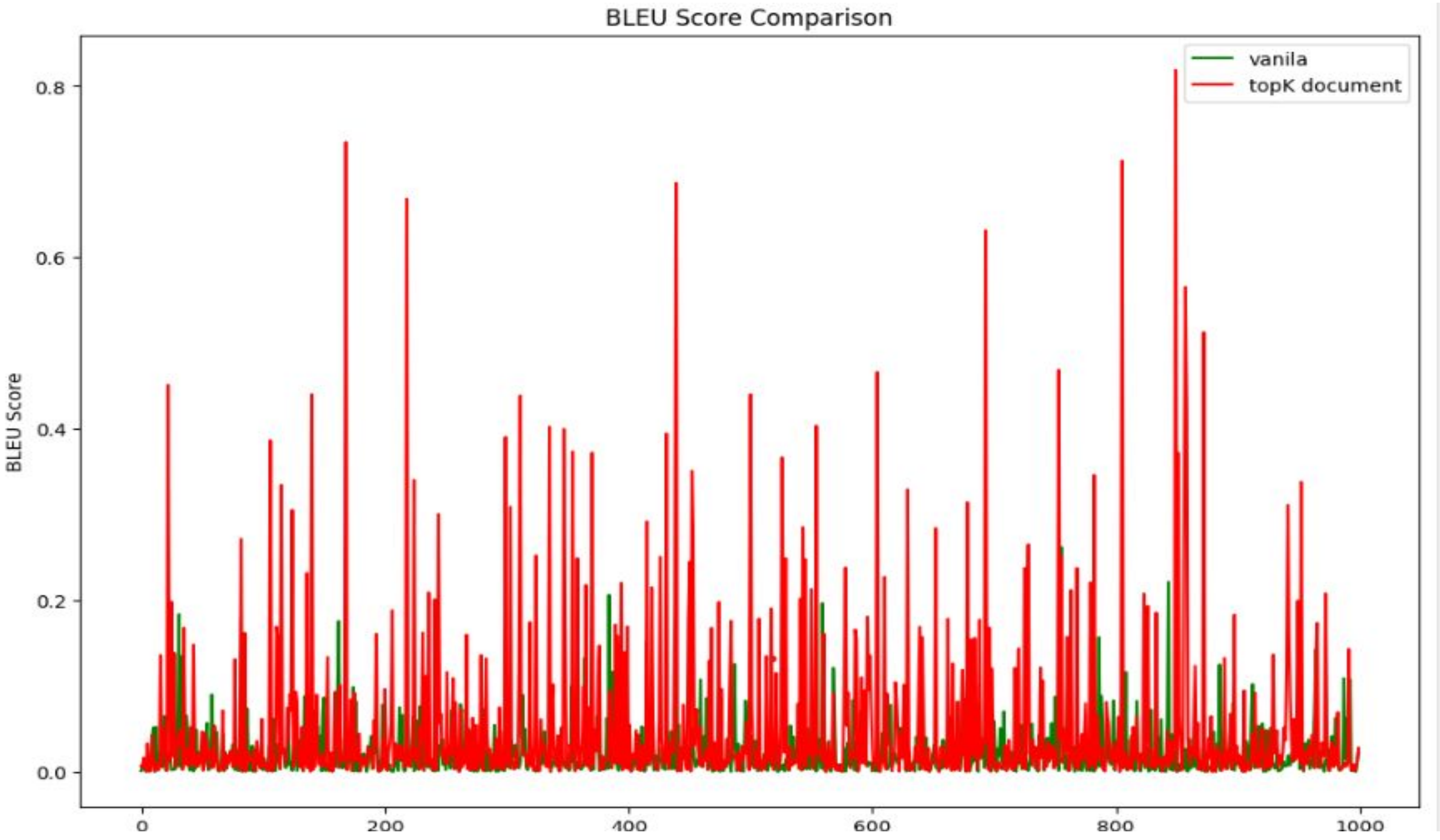
Evaluation Results - Rouge2 Comparison on LLaMa2



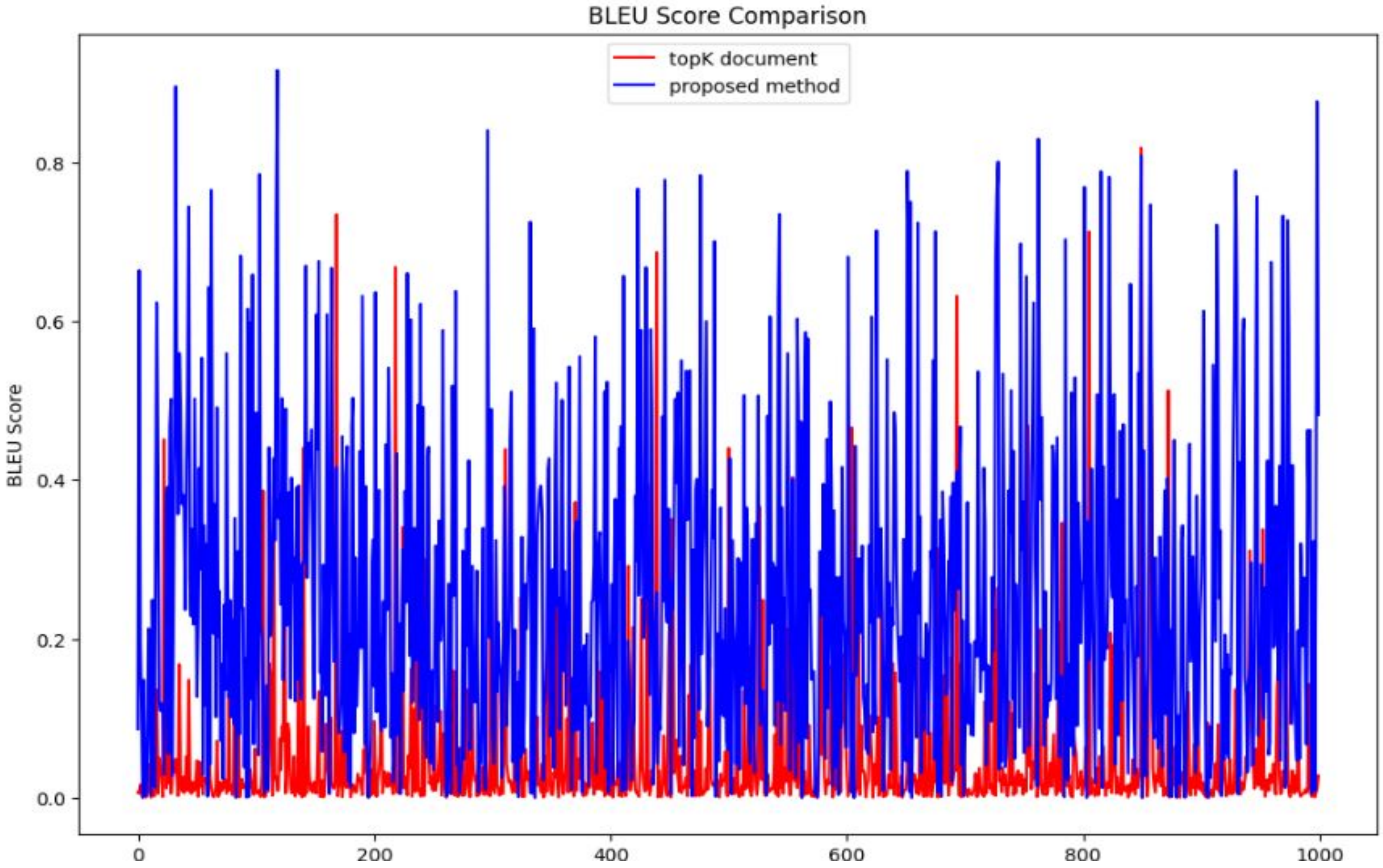
Evaluation Results - RougeL Comparison on LLaMa2



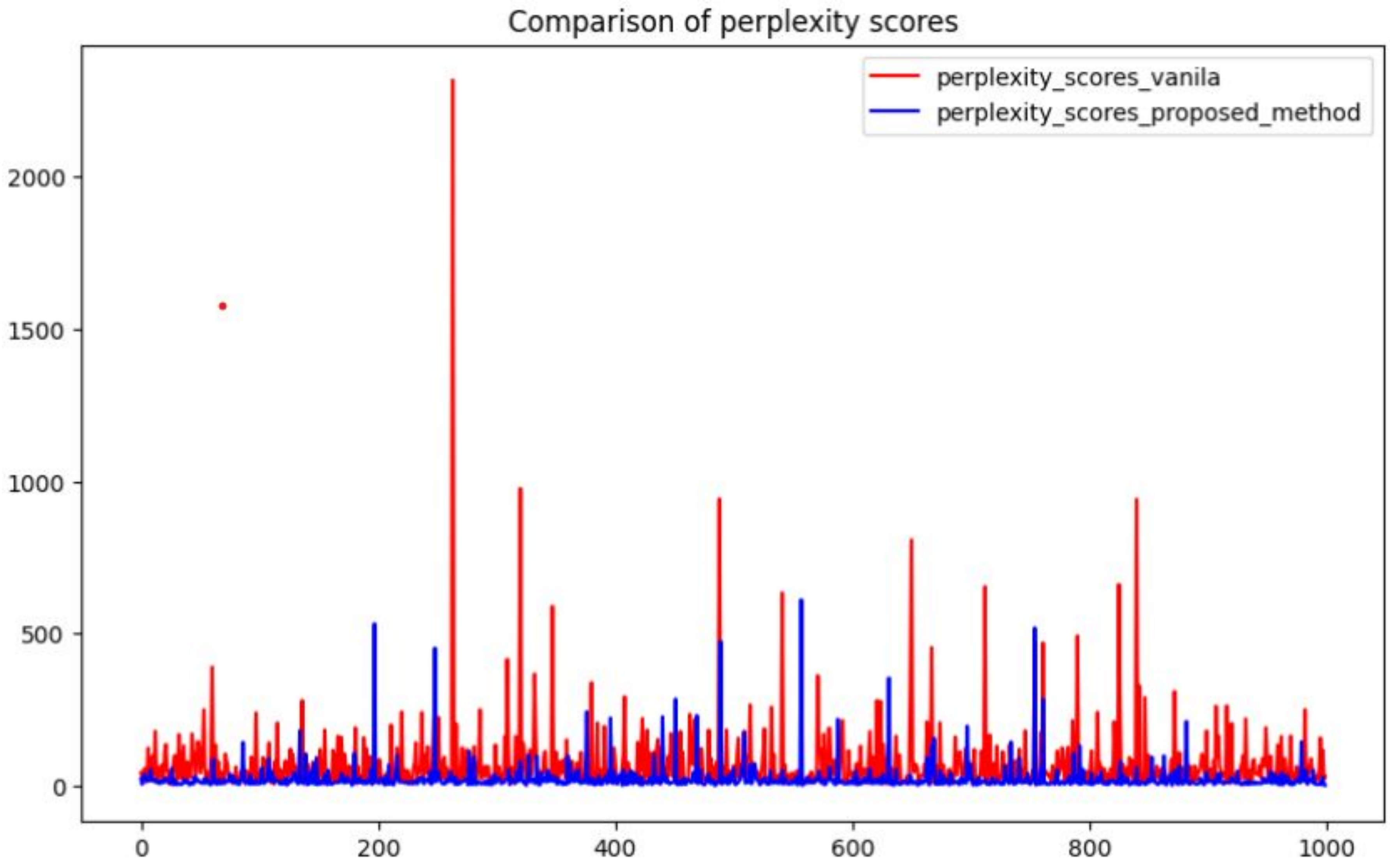
Evaluation Results - BLEU Score Comparison on LLaMa2



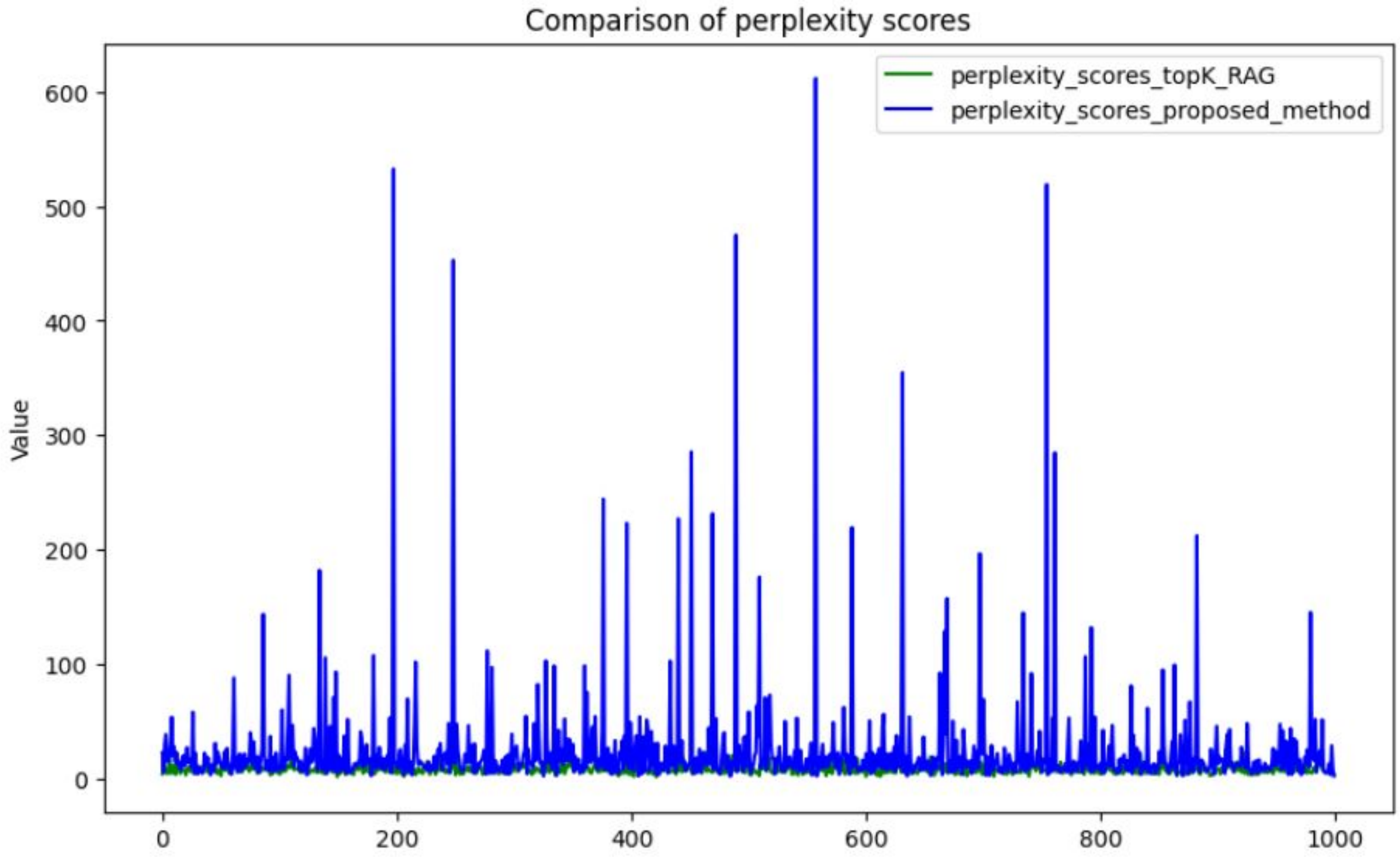
Evaluation Results - BLEU Score Comparison on LLaMa2



Evaluation Results - Perplexity Score Comparison on LLaMa2



Evaluation Results - Perplexity Score Comparison on LLaMa2



Difficulties and Limitations

(1)We don’t have enough GPU to run the inference using full document as context

(2)It takes long time to generate the Wikipedia embeddings due to large size of wikipedia documents

(3)The context size of GPT2 is limited and we can not fit the entire top document into the prompt of GPT2; LLaMa model has long context windows, but it takes GPUs to run experiments and we do not have enough resource to run large scale experiments compared to industry

(4)On the evaluation of truthfulness, usefulness and trustworthiness, we lack the man power to evaluate the all the outputs (ROUGE, BLUE and Perplexity do not guarantee truthfulness, usefulness and trustworthiness)

(5)The NQdataset evaluation might still not be comprehensive to encompass a wide range of human knowledge

Summary of Contribution

(1)Proposed method on better extracting relevant information for RAG system

(2)Extensive evaluation on proposed method on open source model including GPT2 and LLaMa2; And the proposed method achieve better performance in ROUGE, BLUE and Perplexity compared to two baselines

Future Work

- Evaluation reveals that while LLMs exhibit a certain degree of noise robustness
- Still struggle significantly in terms of negative rejection, information integration, and dealing with false information

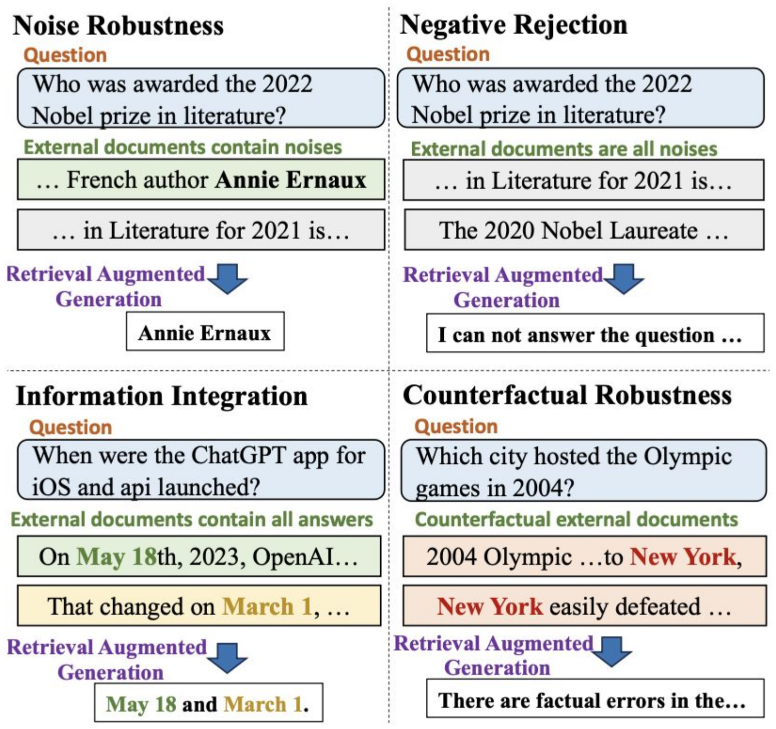


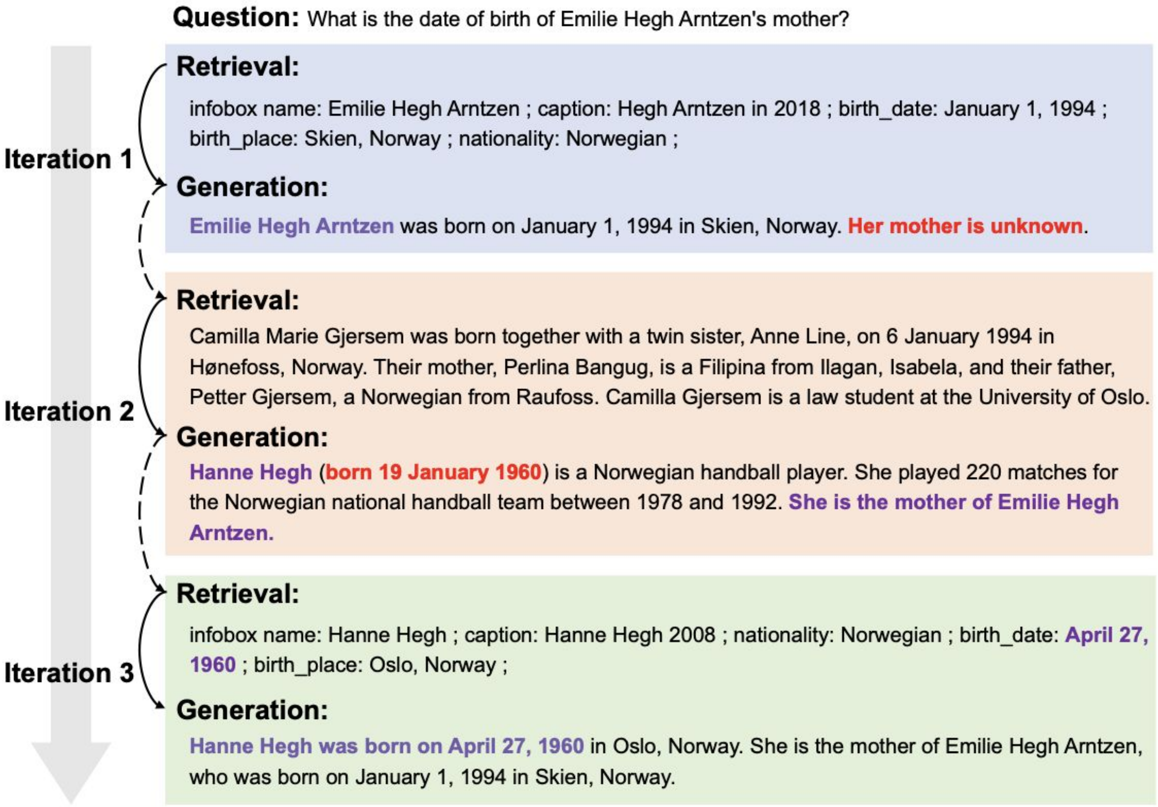
Figure 1: Illustration of 4 kinds of abilities required for retrieval-augmented generation of LLMs.

Chen, J., Lin, H., Han, X., & Sun, L. (2024). Benchmarking Large Language Models in Retrieval-Augmented Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(16), 17754-17762. https://doi.org/10.1609/aaai.v38i16.29728

Future Work

- Proposes terative retrieval-generation collaborative framework

- Able to leverage both parametric and non-parametric knowledge, but also helps to find the correct reasoning path through retrieval-generation interactions



Z. Feng, X. Feng, D. Zhao, M. Yang and B. Qin, "Retrieval-Generation Synergy Augmented Large Language Models," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 11661-11665, doi: 10.1109/ICASSP48485.2024.10448015.