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
- Background:
  - Embedding
  - Vector Database
  - Information Retrieval
  - Large Language Model (LLM)
- Methodology
  - Re-ranking
- Experiment Setup
  - Benchmarks
  - Datasets
  - Metrics,
- Experimental Results and Insights

Background

Retrieval-Augmented Generation (RAG):

- A technique that combines the strengths of retrieval-based and generation-based models.

Benefits:

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

Background

Limitations:

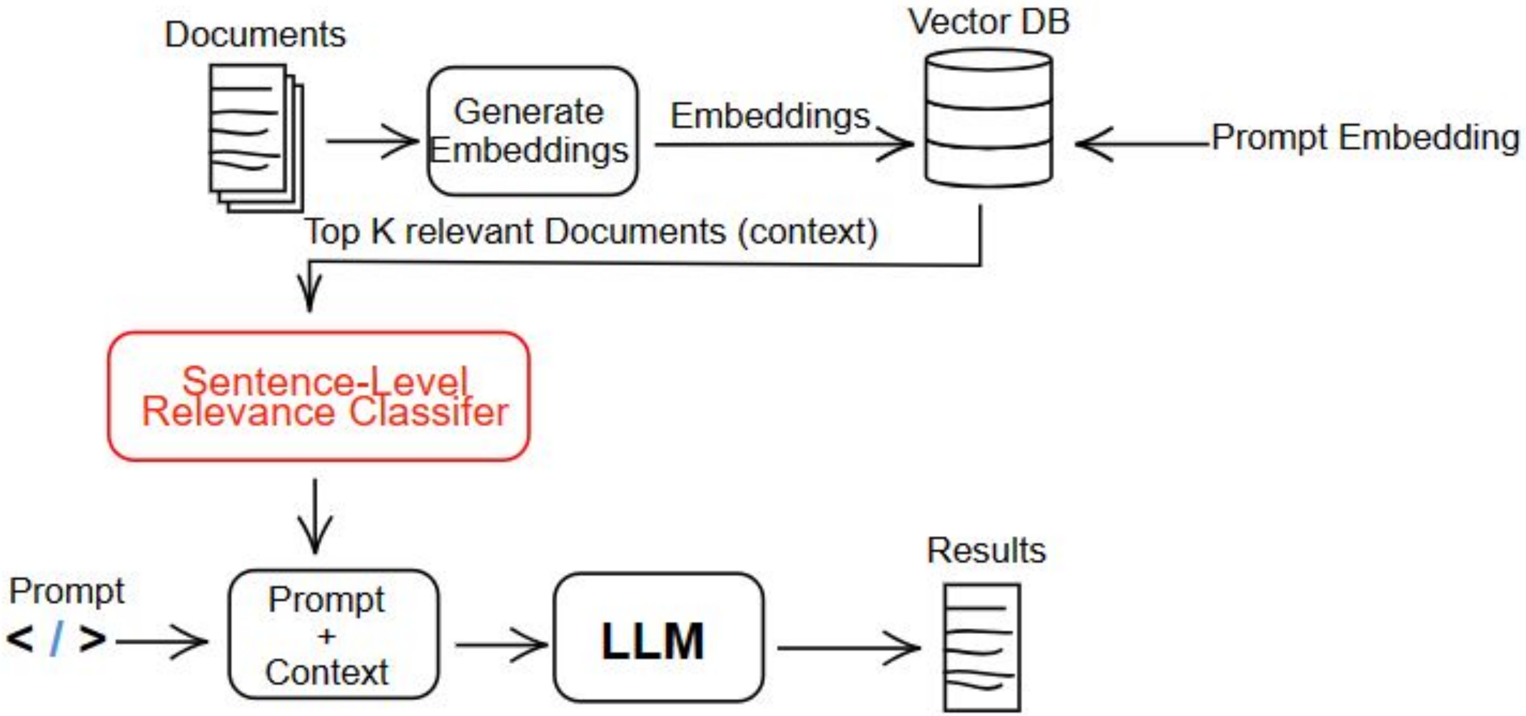
- LLM relies on highly relevant and specific information.
- But traditional methods don’t provide relevant and effective information
- Traditional methods provides long top k documents which are too long to be effective
- Truncation methods are not reliable enough

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



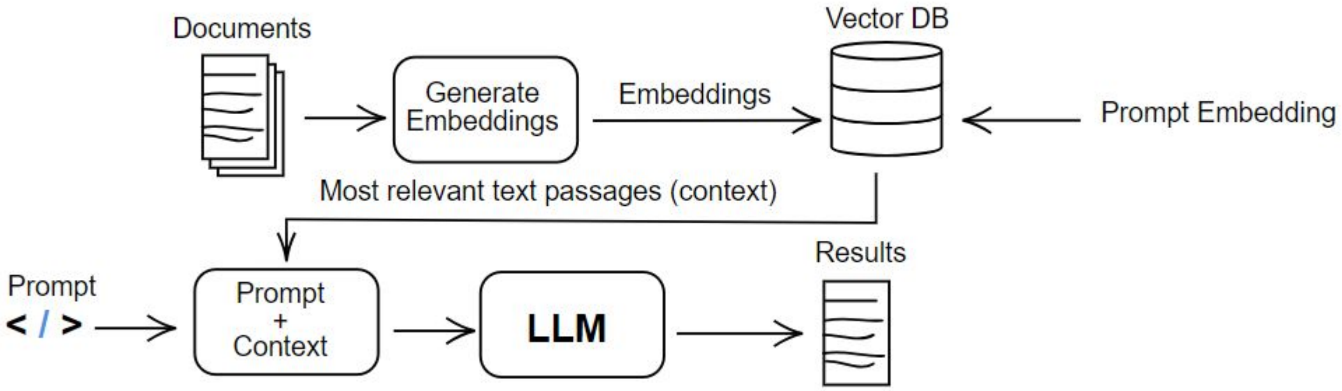
Related Work

- 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.

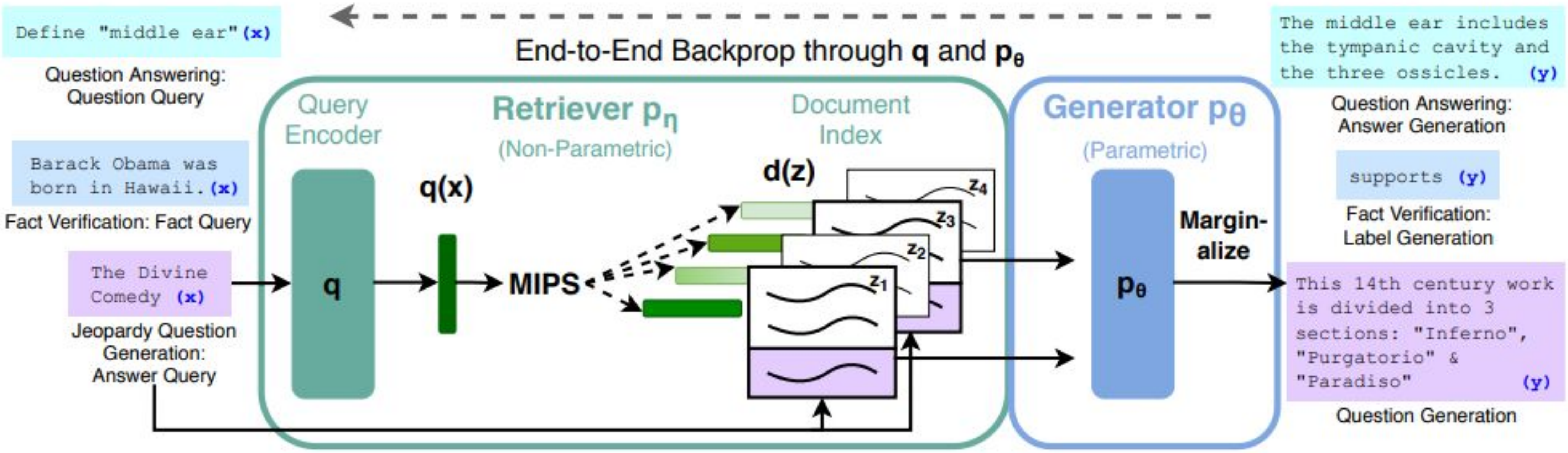
Overview

- 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/

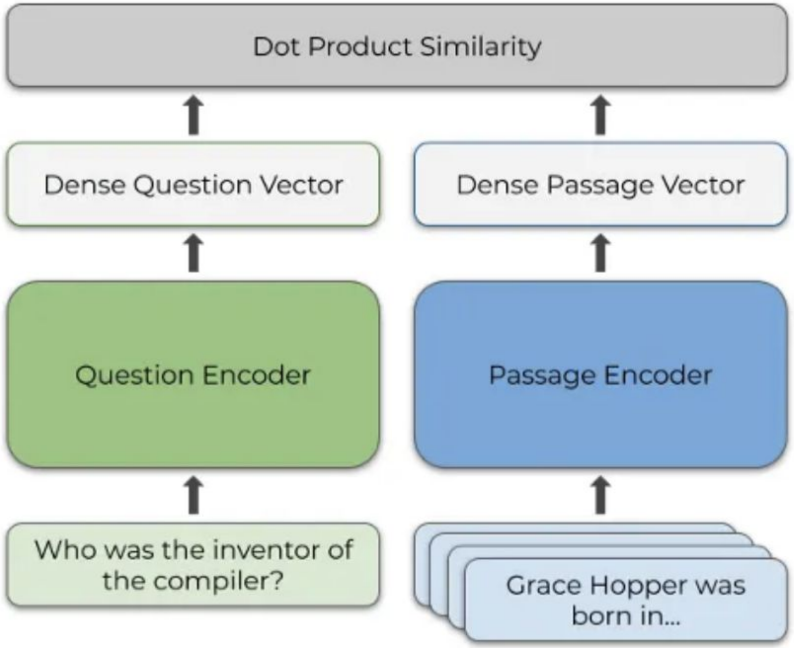
RAG



Define "middle ear" (x)
Question Answering: Question Query
Barack Obama was born in Hawaii. (x)
Fact Verification: Fact Query
The Divine Comedy (x)
Jeopardy Question Generation: Answer Query
End-to-End Backprop through q and p_θ
Query Encoder
q
q(x)
Retriever p_η
(Non-Parametric)
MIPS
Document Index
d(z)
z_4
z_3
z_2
z_1
Generator p_θ
(Parametric)
p_θ
Margin-alize
The middle ear includes the tympanic cavity and the three ossicles. (y)
Question Answering: Answer Generation
supports (y)
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (y)
Question Generation

arXiv:2005.11401

RAG Retriever



Dot Product Similarity

Dense Question Vector

Dense Passage Vector

Question Encoder

Passage Encoder

Who was the inventor of the compiler?

Grace Hopper was born in...

Dense Passage Retrieval (DPR) architecture.

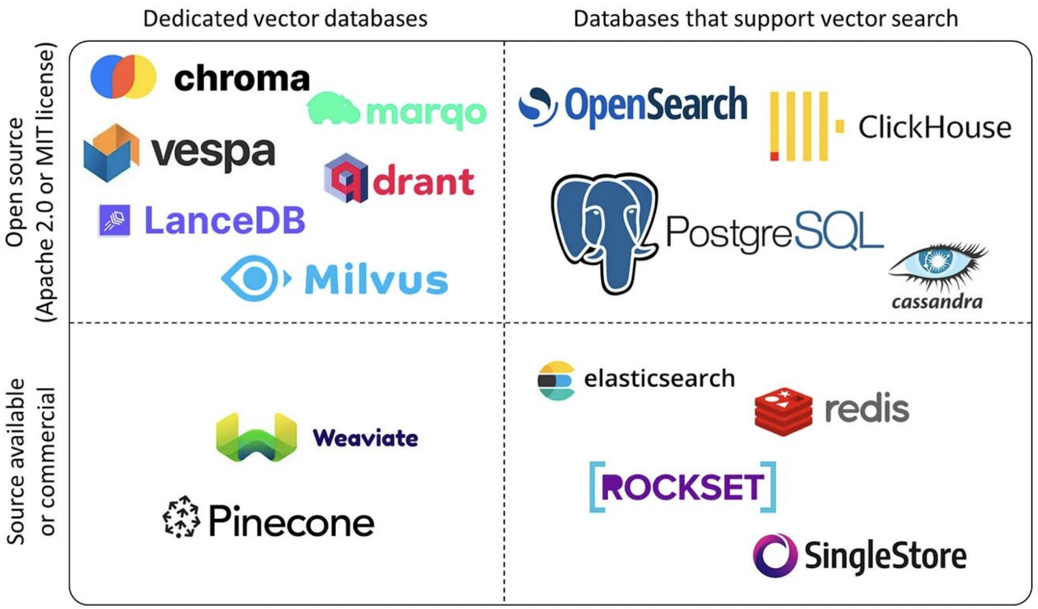
$$p_\eta(z|x) \propto \exp\left(\mathbf{d}(z)^\top \mathbf{q}(x)\right)$$

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

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

Storage

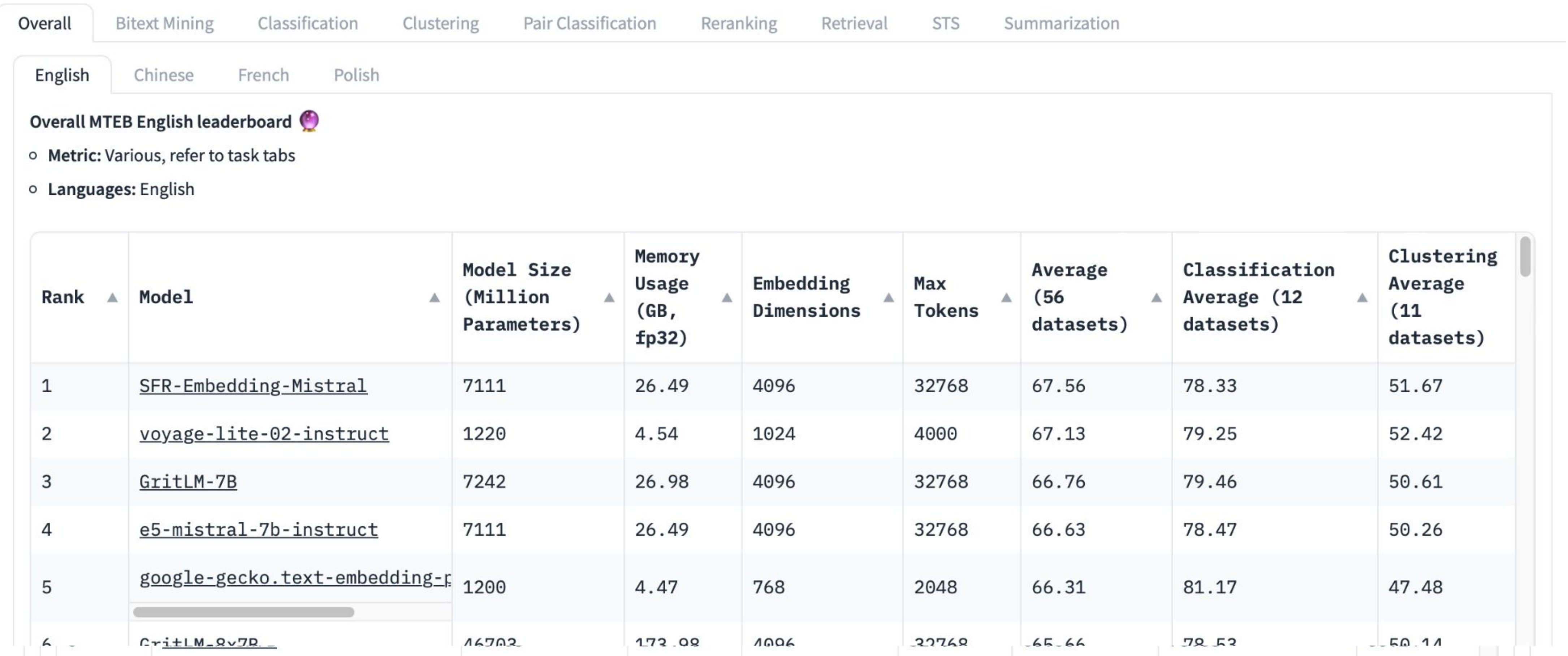
- Vector index (ex. Faiss) and vector database options
- Used Wikipedia dataset


Dedicated vector databases
Databases that support vector search
Open source (Apache 2.0 or MIT license)
Source available or commercial
chroma
marqo
vespa
qdrant
LanceDB
Milvus
OpenSearch
ClickHouse
PostgreSQL
cassandra
Weaviate
Pinecone
elasticsearch
redis
ROCKSET
SingleStore

The landscape of vector databases.

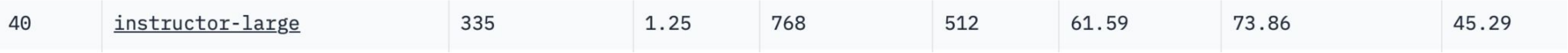
https://blog.det.life/why-you-shouldnt-invest-in-vector-databases-c0cd3f59d23c

Embedding Models



Overall MTEB English leaderboard 🔮

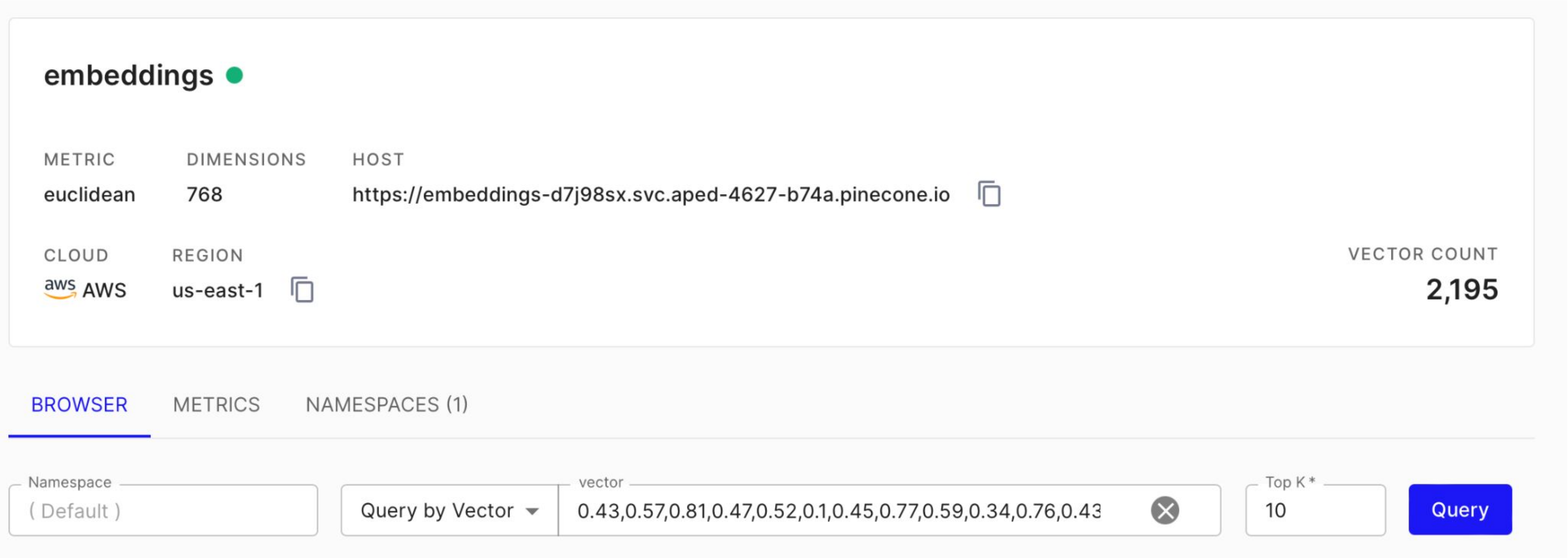
- **Metric:** Various, refer to task tabs
- **Languages:** English



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

Storage



embeddings ●

METRIC: euclidean

DIMENSIONS: 768

HOST: https://embeddings-d7j98sx.svc.aped-4627-b74a.pinecone.io

CLOUD: aws AWS

REGION: us-east-1

VECTOR COUNT: 2,195

Namespace: (Default)

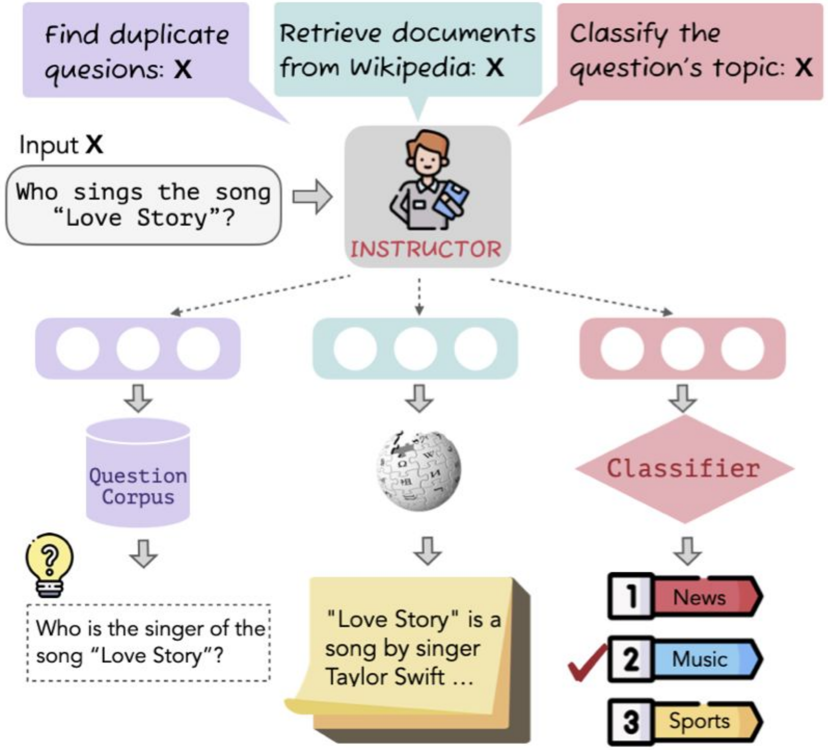
Query by Vector

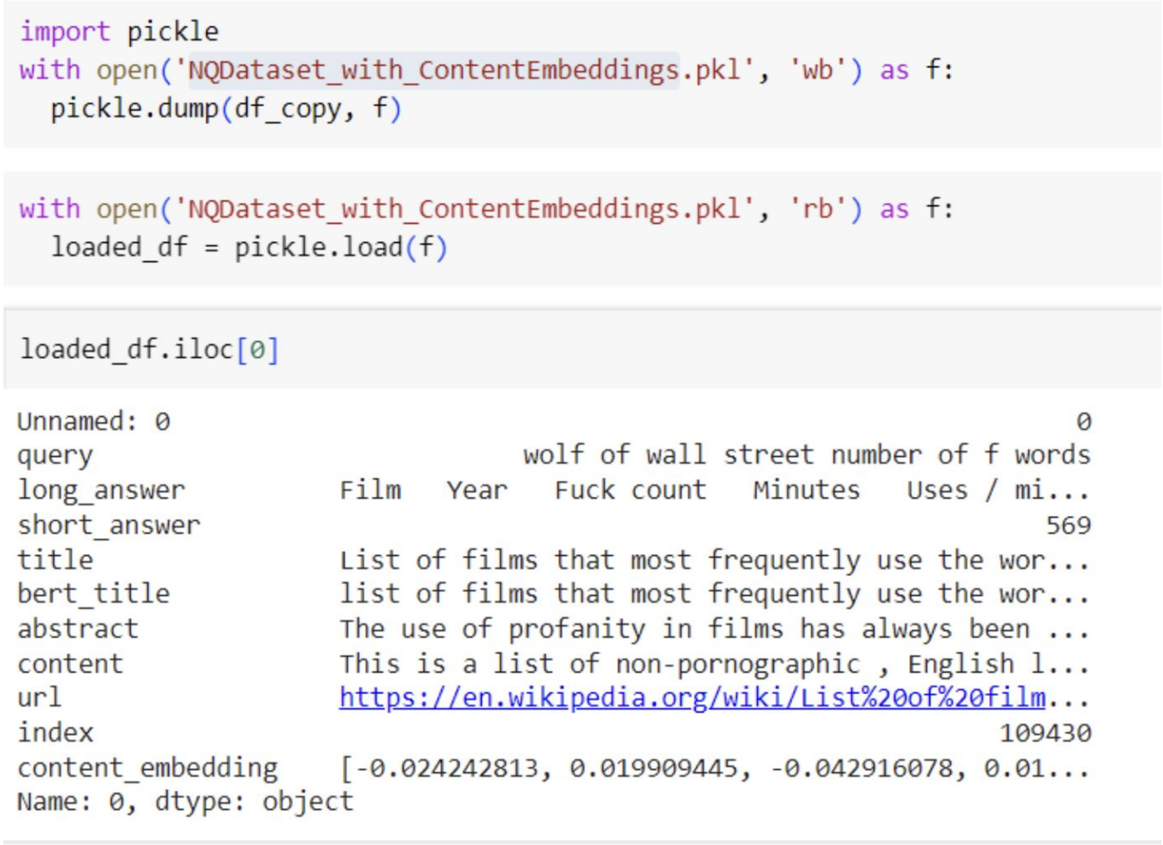
vector: 0.43,0.57,0.81,0.47,0.52,0.1,0.45,0.77,0.59,0.34,0.76,0.43

Top K *: 10

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

Language Model and Information Retrieval


Find duplicate quesions: X
Retrieve documents from Wikipedia: X
Classify the question's topic: X
Input X
Who sings the song "Love Story"?
INSTRUCTOR
Question Corpus
Classifier
Who is the singer of the song "Love Story"?
"Love Story" is a song by singer Taylor Swift ...
1 News
2 Music
3 Sports



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

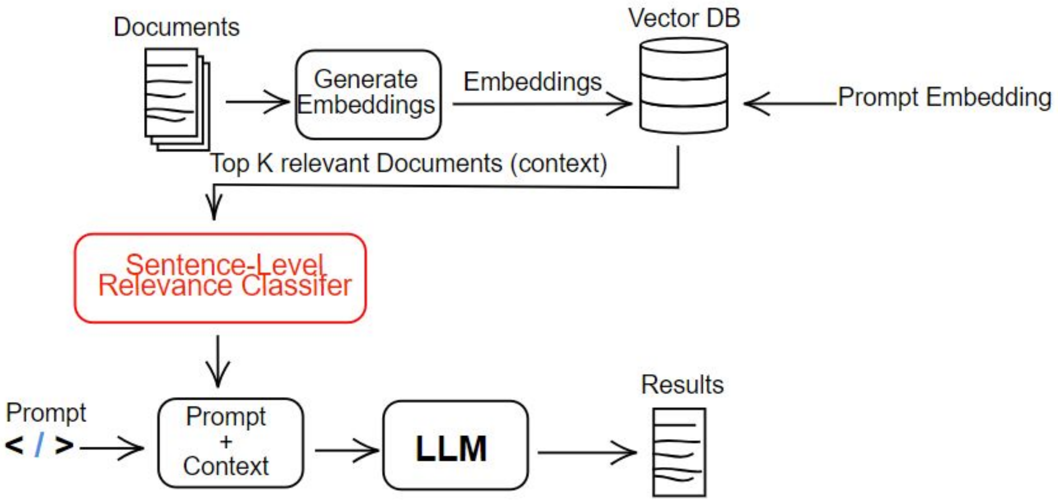
```
with open('NQDataset_with_ContentEmbeddings.pkl', 'rb') as f:
  loaded_df = pickle.load(f)
```

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

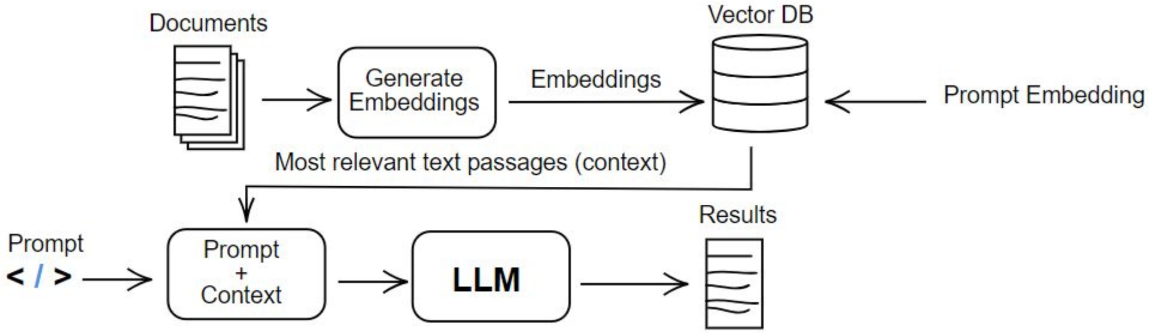
```
Unnamed: 0                                                         0
query                                 wolf of wall street number of f words
long_answer          Film   Year   Fuck count   Minutes   Uses / mi...
short_answer                                                     569
title                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

Proposed vs Existing Framework



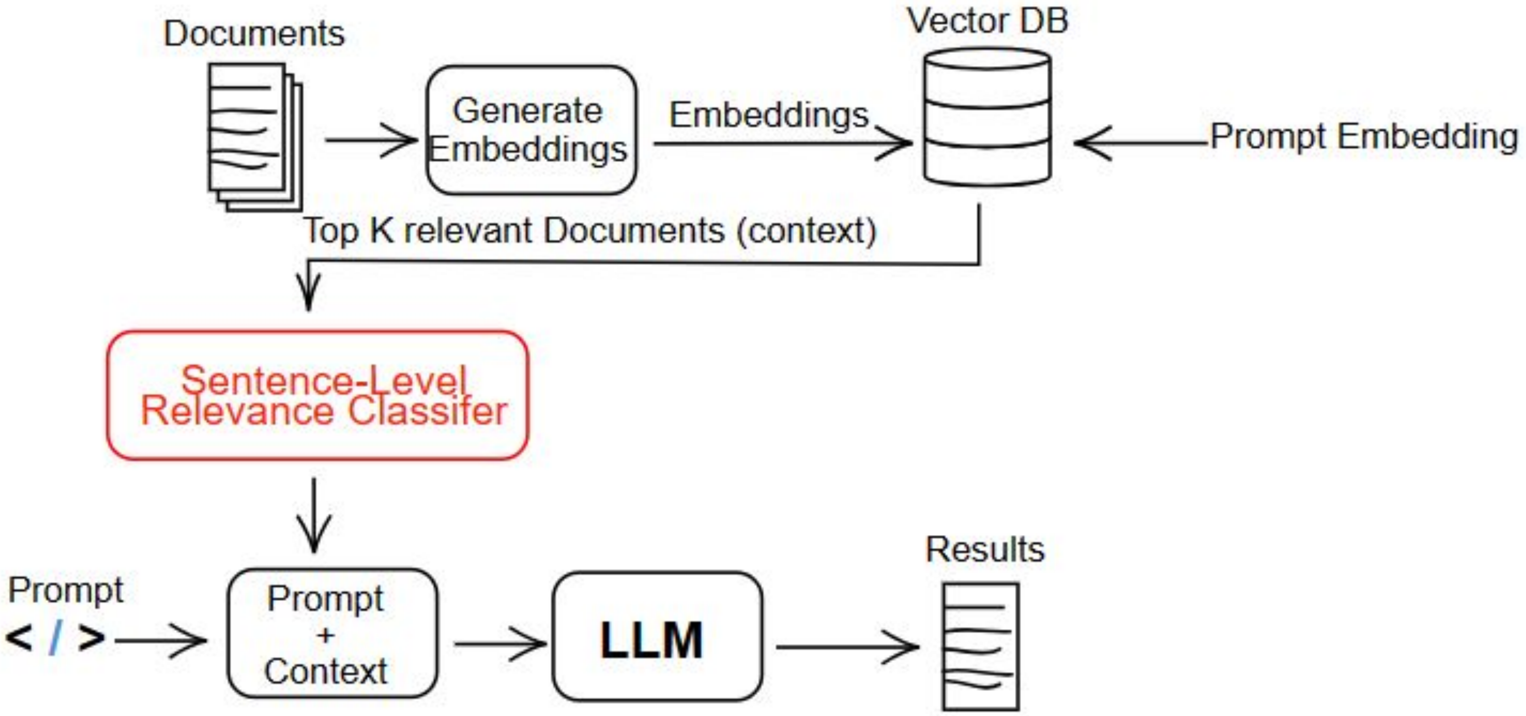
Our Method



Existing Framework

Methodology

- Framework:



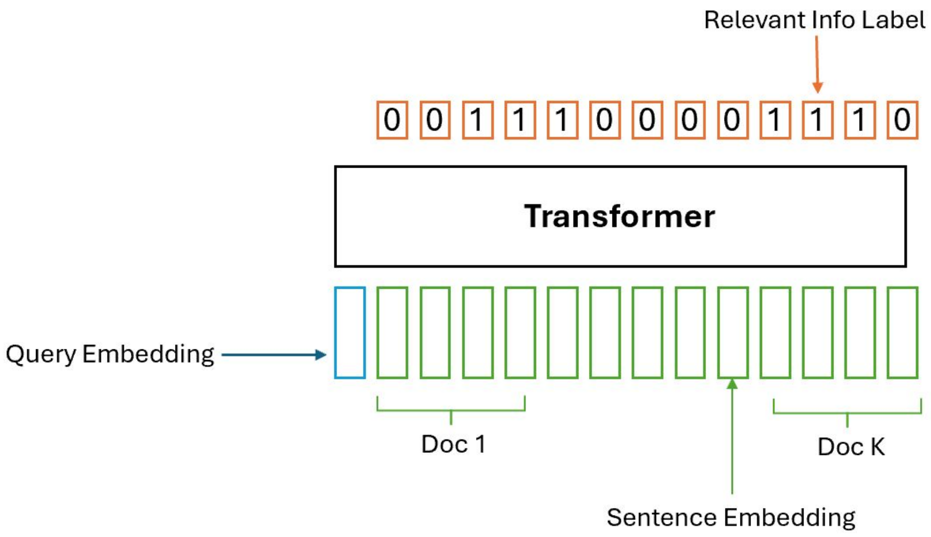
Methodology

- 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.

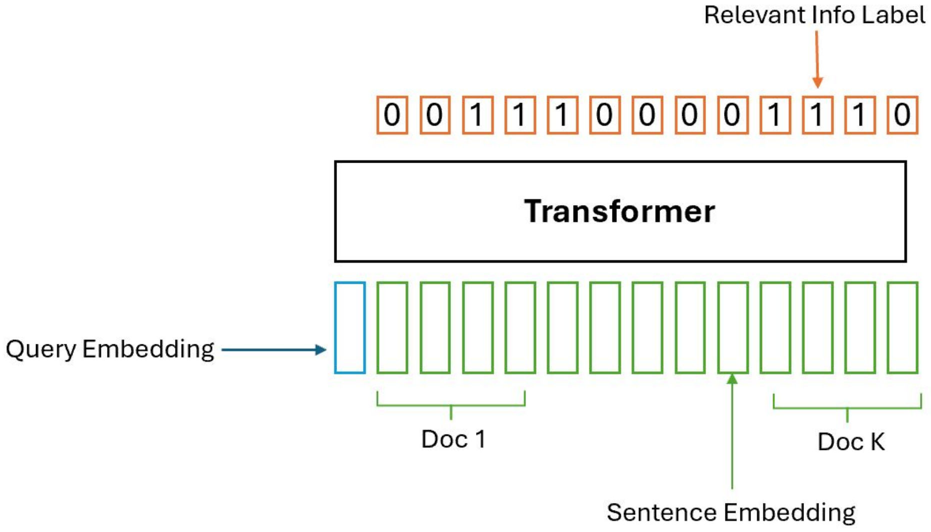
Proposed Method

- 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.

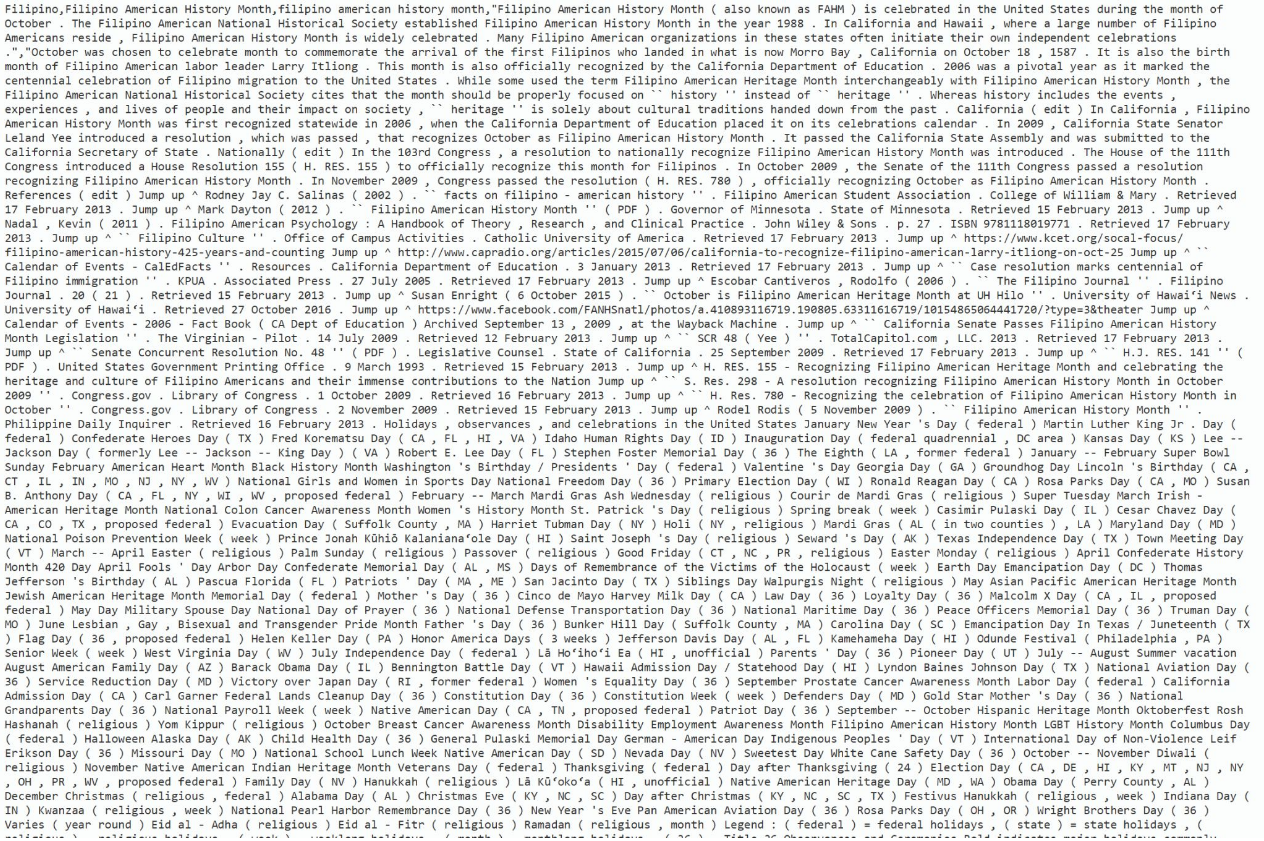


Proposed Method

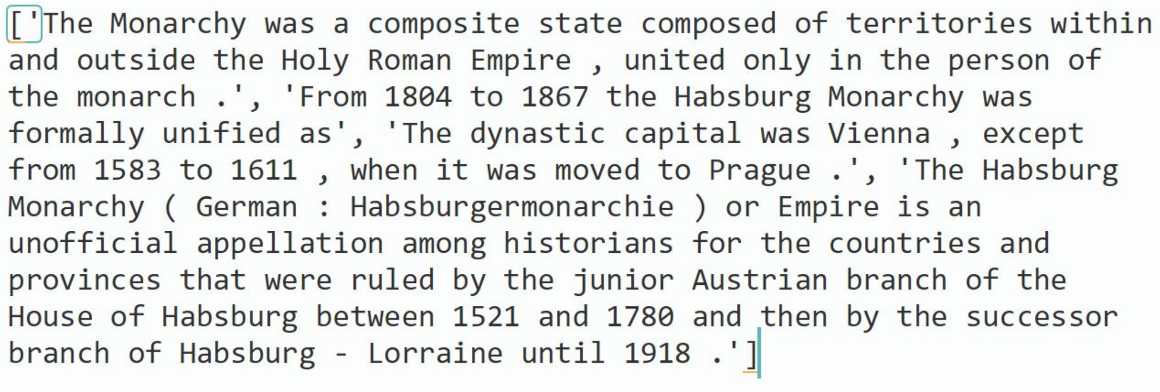
- Transformer: A transformer encoder
- Training:
    - 1335 samples
- Inference:
    - 1000 samples



Top K Doc vs Relevant Sentences



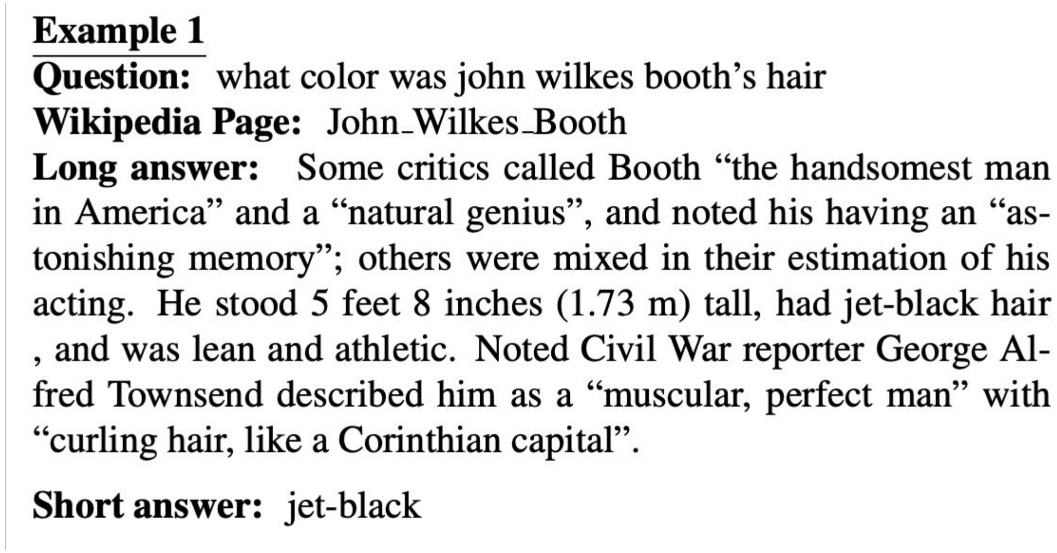
```
Filipino,Filipino American History Month,filipino american history month,"Filipino American History Month ( also known as FAHM ) is celebrated in the United States during the month of
October . The Filipino American National Historical Society established Filipino American History Month in the year 1988 . In California and Hawaii , where a large number of Filipino
Americans reside , Filipino American History Month is widely celebrated . Many Filipino American organizations in these states often initiate their own independent celebrations
.","October was chosen to celebrate month to commemorate the arrival of the first Filipinos who landed in what is now Morro Bay , California on October 18 , 1587 . It is also the birth
month of Filipino American labor leader Larry Itliong . This month is also officially recognized by the California Department of Education . 2006 was a pivotal year as it marked the
centennial celebration of Filipino migration to the United States . While some used the term Filipino American Heritage Month interchangeably with Filipino American History Month , the
Filipino American National Historical Society cites that the month should be properly focused on `` history '' instead of `` heritage '' . Whereas history includes the events ,
experiences , and lives of people and their impact on society , `` heritage '' is solely about cultural traditions handed down from the past . California ( edit ) In California , Filipino
American History Month was first recognized statewide in 2006 , when the California Department of Education placed it on its celebrations calendar . In 2009 , California State Senator
Leland Yee introduced a resolution , which was passed , that recognizes October as Filipino American History Month . It passed the California State Assembly and was submitted to the
California Secretary of State . Nationally ( edit ) In the 103rd Congress , a resolution to nationally recognize Filipino American History Month was introduced . The House of the 111th
Congress introduced a House Resolution 155 ( H. RES. 155 ) to officially recognize this month for Filipinos . In October 2009 , the Senate of the 111th Congress passed a resolution
recognizing Filipino American History Month . In November 2009 , Congress passed the resolution ( H. RES. 780 ) , officially recognizing October as Filipino American History Month .
References ( edit ) Jump up ^ Rodney Jay C. Salinas ( 2002 ) . `` facts on filipino - american history '' . Filipino American Student Association . College of William & Mary . Retrieved
17 February 2013 . Jump up ^ Mark Dayton ( 2012 ) . `` Filipino American History Month '' ( PDF ) . Governor of Minnesota . State of Minnesota . Retrieved 15 February 2013 . Jump up ^
Nadal , Kevin ( 2011 ) . Filipino American Psychology : A Handbook of Theory , Research , and Clinical Practice . John Wiley & Sons . p. 27 . ISBN 9781118019771 . Retrieved 17 February
2013 . Jump up ^ `` Filipino Culture '' . Office of Campus Activities . Catholic University of America . Retrieved 17 February 2013 . Jump up ^ https://www.kcet.org/socal-focus/
filipino-american-history-425-years-and-counting Jump up ^ http://www.capradio.org/articles/2015/07/06/california-to-recognize-filipino-american-larry-itliong-on-oct-25 Jump up ^ ``
Calendar of Events - CalEdFacts '' . Resources . California Department of Education . 3 January 2013 . Retrieved 17 February 2013 . Jump up ^ `` Case resolution marks centennial of
Filipino immigration '' . KPUA . Associated Press . 27 July 2005 . Retrieved 17 February 2013 . Jump up ^ Escobar Cantiveros , Rodolfo ( 2006 ) . `` The Filipino Journal '' . Filipino
Journal . 20 ( 21 ) . Retrieved 15 February 2013 . Jump up ^ Susan Enright ( 6 October 2015 ) . `` October is Filipino American Heritage Month at UH Hilo '' . University of Hawaiʻi News .
University of Hawaiʻi . Retrieved 27 October 2016 . Jump up ^ https://www.facebook.com/FANHSnatl/photos/a.410893116719.190805.63311616719/10154865064441720/?type=3&theater Jump up ^
Calendar of Events - 2006 - Fact Book ( CA Dept of Education ) Archived September 13 , 2009 , at the Wayback Machine . Jump up ^ `` California Senate Passes Filipino American History
Month Legislation '' . The Virginian - Pilot . 14 July 2009 . Retrieved 12 February 2013 . Jump up ^ `` SCR 48 ( Yee ) '' . TotalCapitol.com , LLC. 2013 . Retrieved 17 February 2013 .
Jump up ^ `` Senate Concurrent Resolution No. 48 '' ( PDF ) . Legislative Counsel . State of California . 25 September 2009 . Retrieved 17 February 2013 . Jump up ^ `` H.J. RES. 141 '' (
PDF ) . United States Government Printing Office . 9 March 1993 . Retrieved 15 February 2013 . Jump up ^ H. RES. 155 - Recognizing Filipino American Heritage Month and celebrating the
heritage and culture of Filipino Americans and their immense contributions to the Nation Jump up ^ `` S. Res. 298 - A resolution recognizing Filipino American History Month in October
2009 '' . Congress.gov . Library of Congress . 1 October 2009 . Retrieved 16 February 2013 . Jump up ^ `` H. Res. 780 - Recognizing the celebration of Filipino American History Month in
October '' . Congress.gov . Library of Congress . 2 November 2009 . Retrieved 15 February 2013 . Jump up ^ Rodel Rodis ( 5 November 2009 ) . `` Filipino American History Month '' .
Philippine Daily Inquirer . Retrieved 16 February 2013 . Holidays , observances , and celebrations in the United States January New Year 's Day ( federal ) Martin Luther King Jr . Day (
federal ) Confederate Heroes Day ( TX ) Fred Korematsu Day ( CA , FL , HI , VA ) Idaho Human Rights Day ( ID ) Inauguration Day ( federal quadrennial , DC area ) Kansas Day ( KS ) Lee --
Jackson Day ( formerly Lee -- Jackson -- King Day ) ( VA ) Robert E. Lee Day ( FL ) Stephen Foster Memorial Day ( 36 ) The Eighth ( LA , former federal ) January -- February Super Bowl
Sunday February American Heart Month Black History Month Washington 's Birthday / Presidents ' Day ( federal ) Valentine 's Day Georgia Day ( GA ) Groundhog Day Lincoln 's Birthday ( CA ,
CT , IL , IN , MO , NJ , NY , WV ) National Girls and Women in Sports Day National Freedom Day ( 36 ) Primary Election Day ( WI ) Ronald Reagan Day ( CA ) Rosa Parks Day ( CA , MO ) Susan
B. Anthony Day ( CA , FL , NY , WI , WV , proposed federal ) February -- March Mardi Gras Ash Wednesday ( religious ) Courir de Mardi Gras ( religious ) Super Tuesday March Irish -
American Heritage Month National Colon Cancer Awareness Month Women 's History Month St. Patrick 's Day ( religious ) Spring break ( week ) Casimir Pulaski Day ( IL ) Cesar Chavez Day (
CA , CO , TX , proposed federal ) Evacuation Day ( Suffolk County , MA ) Harriet Tubman Day ( NY ) Holi ( NY , religious ) Mardi Gras ( AL ( in two counties ) , LA ) Maryland Day ( MD )
National Poison Prevention Week ( week ) Prince Jonah Kūhiō Kalanianaʻole Day ( HI ) Saint Joseph 's Day ( religious ) Seward 's Day ( AK ) Texas Independence Day ( TX ) Town Meeting Day
( VT ) March -- April Easter ( religious ) Palm Sunday ( religious ) Passover ( religious ) Good Friday ( CT , NC , PR , religious ) Easter Monday ( religious ) April Confederate History
Month 420 Day April Fools ' Day Arbor Day Confederate Memorial Day ( AL , MS ) Days of Remembrance of the Victims of the Holocaust ( week ) Earth Day Emancipation Day ( DC ) Thomas
Jefferson 's Birthday ( AL ) Pascua Florida ( FL ) Patriots ' Day ( MA , ME ) San Jacinto Day ( TX ) Siblings Day Walpurgis Night ( religious ) May Asian Pacific American Heritage Month
Jewish American Heritage Month Memorial Day ( federal ) Mother 's Day ( 36 ) Cinco de Mayo Harvey Milk Day ( CA ) Law Day ( 36 ) Loyalty Day ( 36 ) Malcolm X Day ( CA , IL , proposed
federal ) May Day Military Spouse Day National Day of Prayer ( 36 ) National Defense Transportation Day ( 36 ) National Maritime Day ( 36 ) Peace Officers Memorial Day ( 36 ) Truman Day (
MO ) June Lesbian , Gay , Bisexual and Transgender Pride Month Father 's Day ( 36 ) Bunker Hill Day ( Suffolk County , MA ) Carolina Day ( SC ) Emancipation Day In Texas / Juneteenth ( TX
) Flag Day ( 36 , proposed federal ) Helen Keller Day ( PA ) Honor America Days ( 3 weeks ) Jefferson Davis Day ( AL , FL ) Kamehameha Day ( HI ) Odunde Festival ( Philadelphia , PA )
Senior Week ( week ) West Virginia Day ( WV ) July Independence Day ( federal ) Lā Hoʻihoʻi Ea ( HI , unofficial ) Parents ' Day ( 36 ) Pioneer Day ( UT ) July -- August Summer vacation
August American Family Day ( AZ ) Barack Obama Day ( IL ) Bennington Battle Day ( VT ) Hawaii Admission Day / Statehood Day ( HI ) Lyndon Baines Johnson Day ( TX ) National Aviation Day (
36 ) Service Reduction Day ( MD ) Victory over Japan Day ( RI , former federal ) Women 's Equality Day ( 36 ) September Prostate Cancer Awareness Month Labor Day ( federal ) California
Admission Day ( CA ) Carl Garner Federal Lands Cleanup Day ( 36 ) Constitution Day ( 36 ) Constitution Week ( week ) Defenders Day ( MD ) Gold Star Mother 's Day ( 36 ) National
Grandparents Day ( 36 ) National Payroll Week ( week ) Native American Day ( CA , TN , proposed federal ) Patriot Day ( 36 ) September -- October Hispanic Heritage Month Oktoberfest Rosh
Hashanah ( religious ) Yom Kippur ( religious ) October Breast Cancer Awareness Month Disability Employment Awareness Month Filipino American History Month LGBT History Month Columbus Day
( federal ) Halloween Alaska Day ( AK ) Child Health Day ( 36 ) General Pulaski Memorial Day German - American Day Indigenous Peoples ' Day ( VT ) International Day of Non-Violence Leif
Erikson Day ( 36 ) Missouri Day ( MO ) National School Lunch Week Native American Day ( SD ) Nevada Day ( NV ) Sweetest Day White Cane Safety Day ( 36 ) October -- November Diwali (
religious ) November Native American Indian Heritage Month Veterans Day ( federal ) Thanksgiving ( federal ) Day after Thanksgiving ( 24 ) Election Day ( CA , DE , HI , KY , MT , NJ , NY
, OH , PR , WV , proposed federal ) Family Day ( NV ) Hanukkah ( religious ) Lā Kūʻokoʻa ( HI , unofficial ) Native American Heritage Day ( MD , WA ) Obama Day ( Perry County , AL )
December Christmas ( religious , federal ) Alabama Day ( AL ) Christmas Eve ( KY , NC , SC ) Day after Christmas ( KY , NC , SC , TX ) Festivus Hanukkah ( religious , week ) Indiana Day (
IN ) Kwanzaa ( religious , week ) National Pearl Harbor Remembrance Day ( 36 ) New Year 's Eve Pan American Aviation Day ( 36 ) Rosa Parks Day ( OH , OR ) Wright Brothers Day ( 36 )
Varies ( year round ) Eid al - Adha ( religious ) Eid al - Fitr ( religious ) Ramadan ( religious , month ) Legend : ( federal ) = federal holidays , ( state ) = state holidays , (
religious ) = religious holidays , ( week ) = weeklong holidays , ( month ) = monthlong holidays , ( 36 ) = Title 36 Observances and Ceremonies Bold indicates major holidays commonly
```



```
['The Monarchy was a composite state composed of territories within
and outside the Holy Roman Empire , united only in the person of
the monarch .', 'From 1804 to 1867 the Habsburg Monarchy was
formally unified as', 'The dynastic capital was Vienna , except
from 1583 to 1611 , when it was moved to Prague .', 'The Habsburg
Monarchy ( German : Habsburgermonarchie ) or Empire is an
unofficial appellation among historians for the countries and
provinces that were ruled by the junior Austrian branch of the
House of Habsburg between 1521 and 1780 and then by the successor
branch of Habsburg - Lorraine until 1918 .']
```

Evaluation

Natural Questions Benchmark



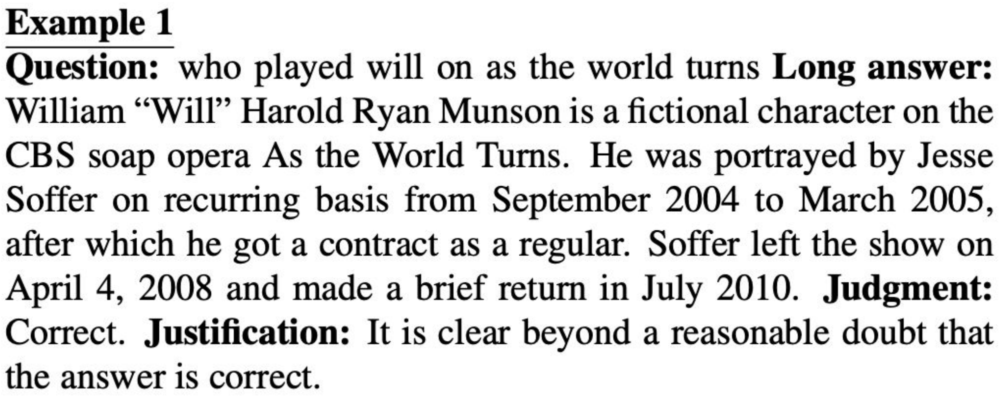
Example 1

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

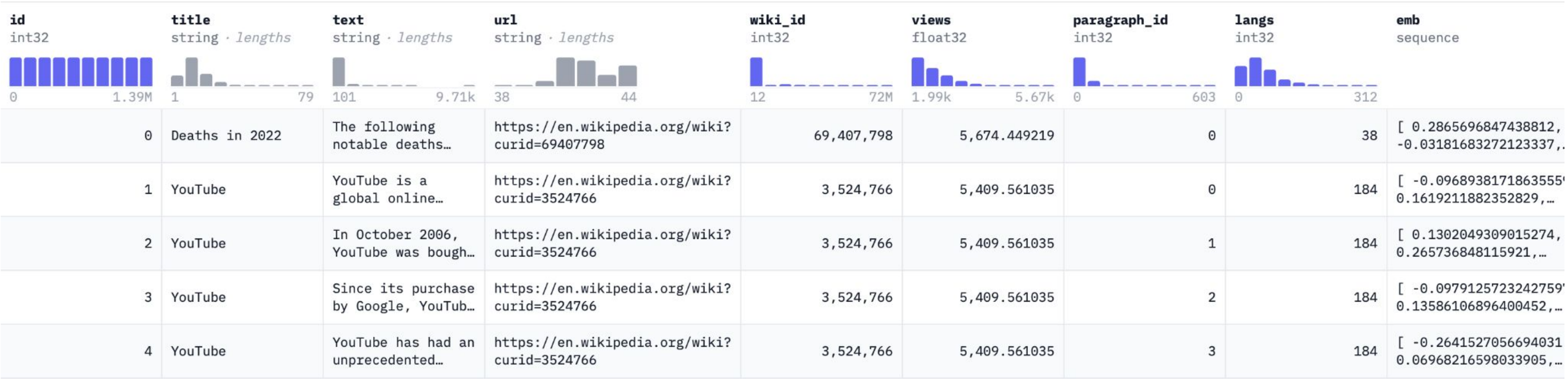


Example 1

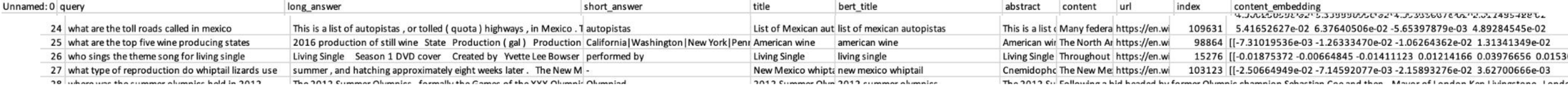
Question: who played will on as the world turns **Long answer:** William "Will" Harold Ryan Munson is a fictional character on the CBS soap opera As the World Turns. He was portrayed by Jesse Soffer on recurring basis from September 2004 to March 2005, after which he got a contract as a regular. Soffer 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

NQ Dataset



id	title	text	url	wiki_id	views	paragraph_id	langs	emb
int32	string · *lengths*	string · *lengths*	string · *lengths*	int32	float32	int32	int32	sequence
0 – 1.39M	1 – 79	101 – 9.71k	38 – 44	12 – 72M	1.99k – 5.67k	0 – 603	0 – 312	
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, -0.0318168327212337,.
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' 0.1358610689640045 2,…
4	YouTube	YouTube has had an unprecedented…	https://en.wikipedia.org/wiki?curid=3524766	3,524,766	5,409.561035	3	184	[-0.2641527056694031 0.0696821659803390 5,…



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 (	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 Olympic	Olympiad	2012 Summer Olym	2012 summer olympics	The 2012 Su	Following a bid headed by former Olympic champion Sebastian Coe and then - Mayor of London Ken Livingstone , Londo			

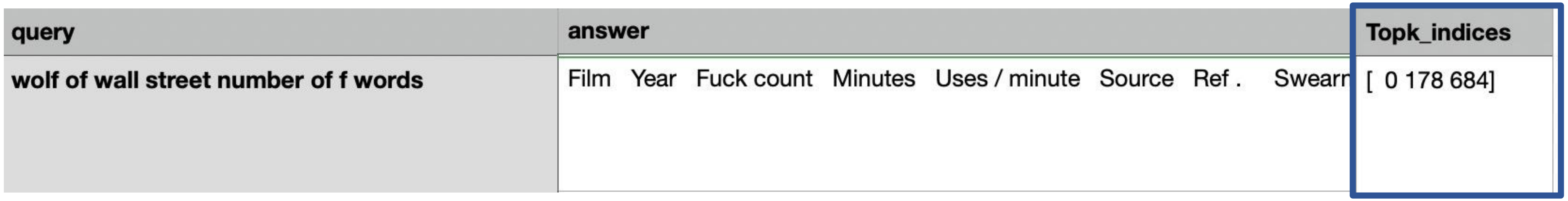
https://huggingface.co/datasets/Cohere/wikipedia-22-12-en-embeddings/viewer/default/train

Evaluation Dataset

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-k document is added for each query&answer pairs.



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

Example of our customized dataset

Evaluation Benchmarks

We choose the GPT-2 and LLaMa as the benchmarks and compare the performance of the following five settings:

1. GPT2 models
2. GPT2 models + top-k documents (as prompts)
3. GPT2 models + selected sentences (as prompts)
4. LLaMa-2 model
5. LLaMa-2 model + top-k documents (as prompts)
6. LLaMa-2 model + selected documents (as prompts)

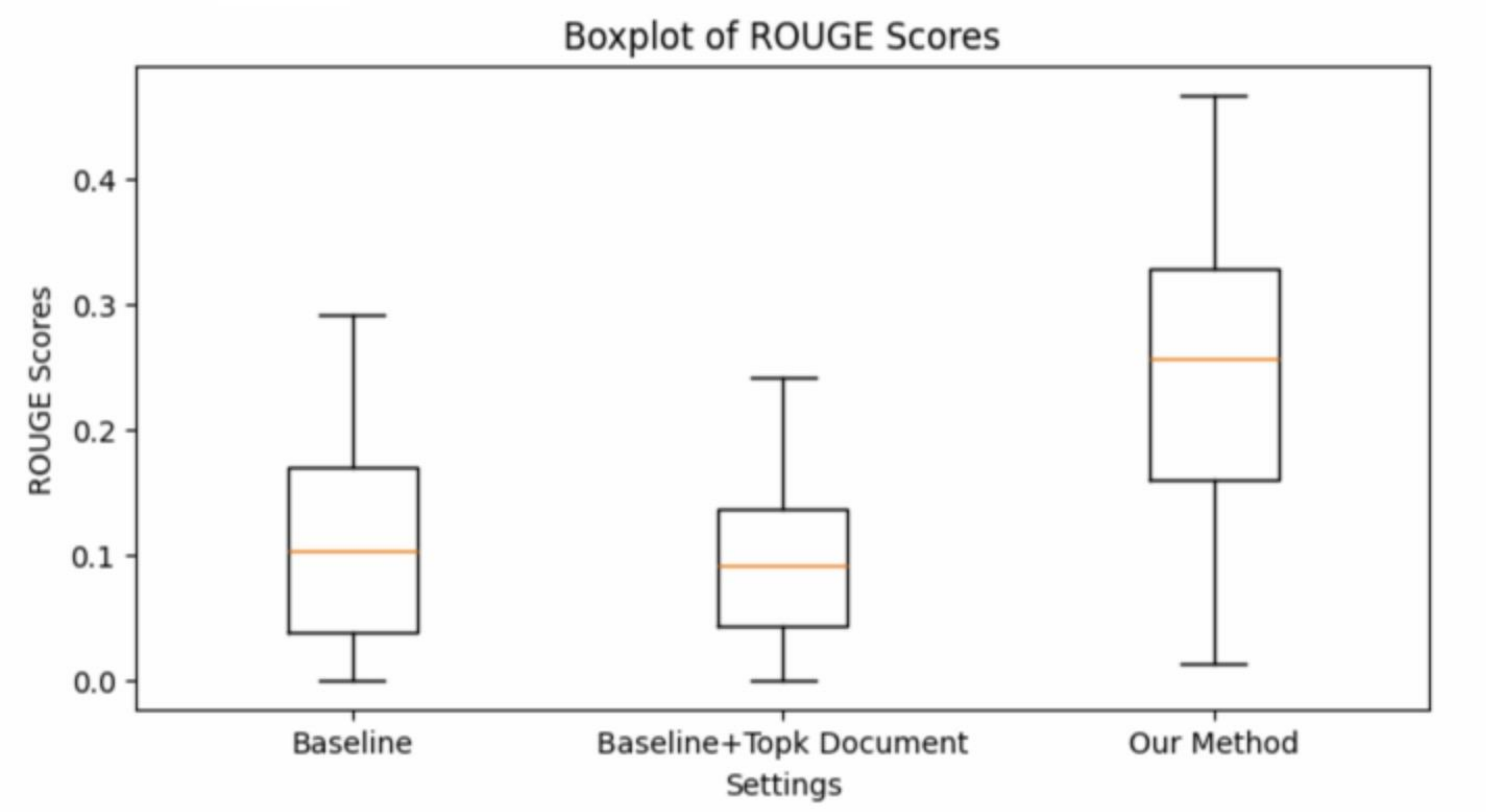
Evaluation Metrics

We choose three metrics to measure the similarity of the machine-translated text to a set of high quality reference translations.

- BLEU score: measure the precision.
- ROUGE score: the harmonic mean of recall and precision.
- Perplexity score: a measure of uncertainty in the value of a sample from a discrete probability distribution.

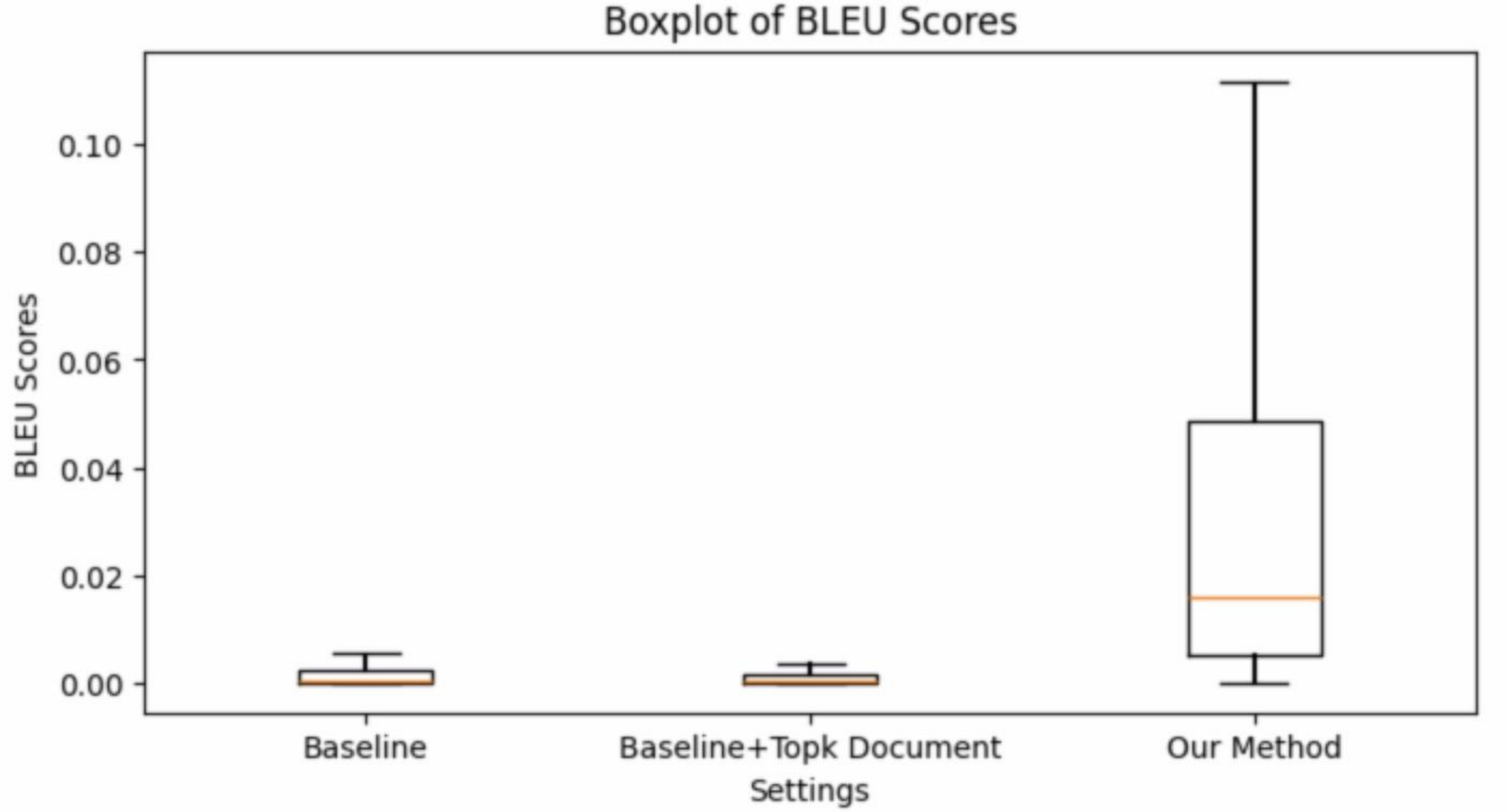
Evaluation Results - ROUGE Comparison of GPT2

- TopK documents mislead the model with portion of irrelevant information.
- Selected sentences provides more accurate context and higher ROUGE score.



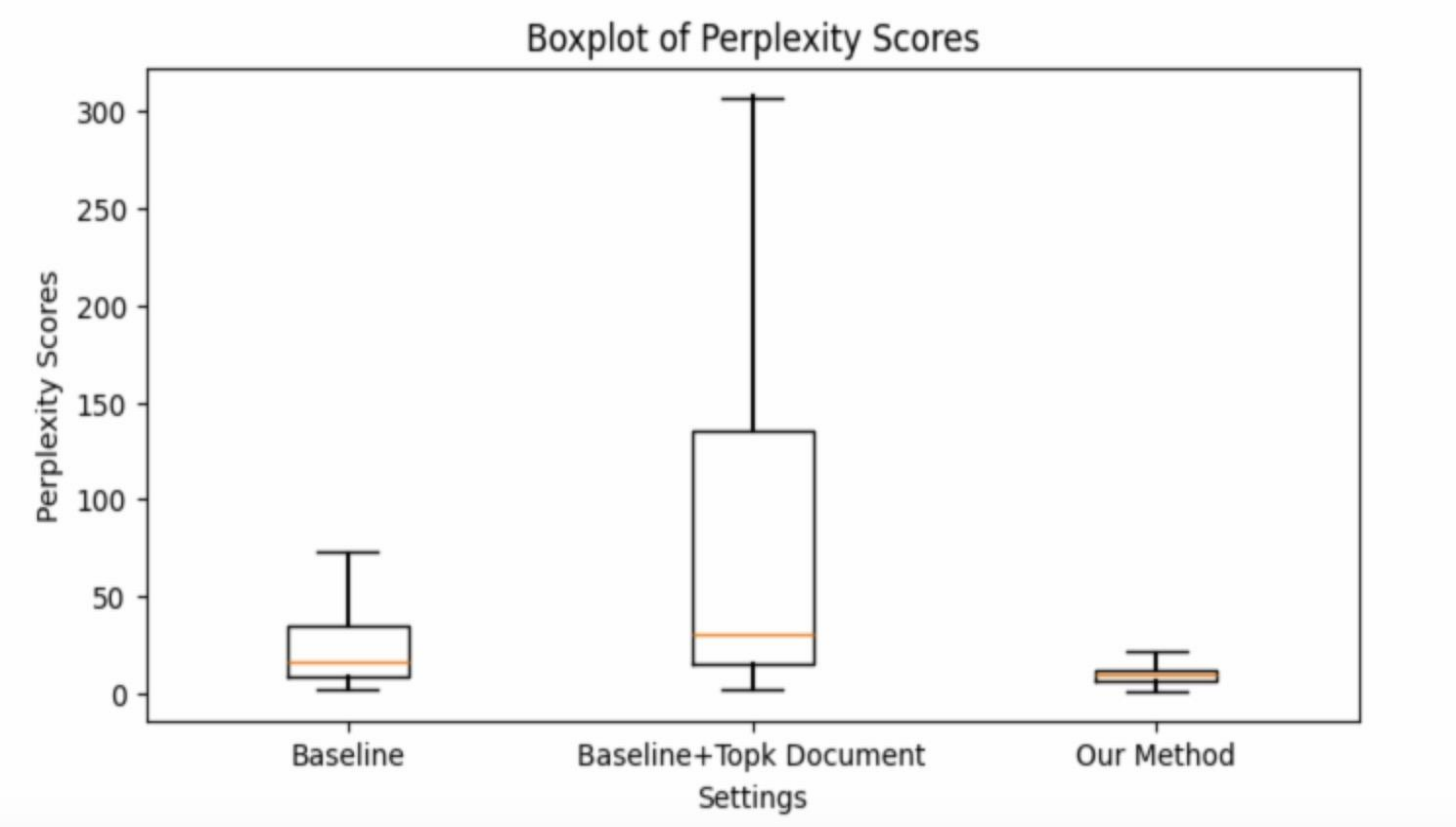
Evaluation Results - BLEU Comparison of GPT2

- BLEU score is greatly increased using our proposed method.



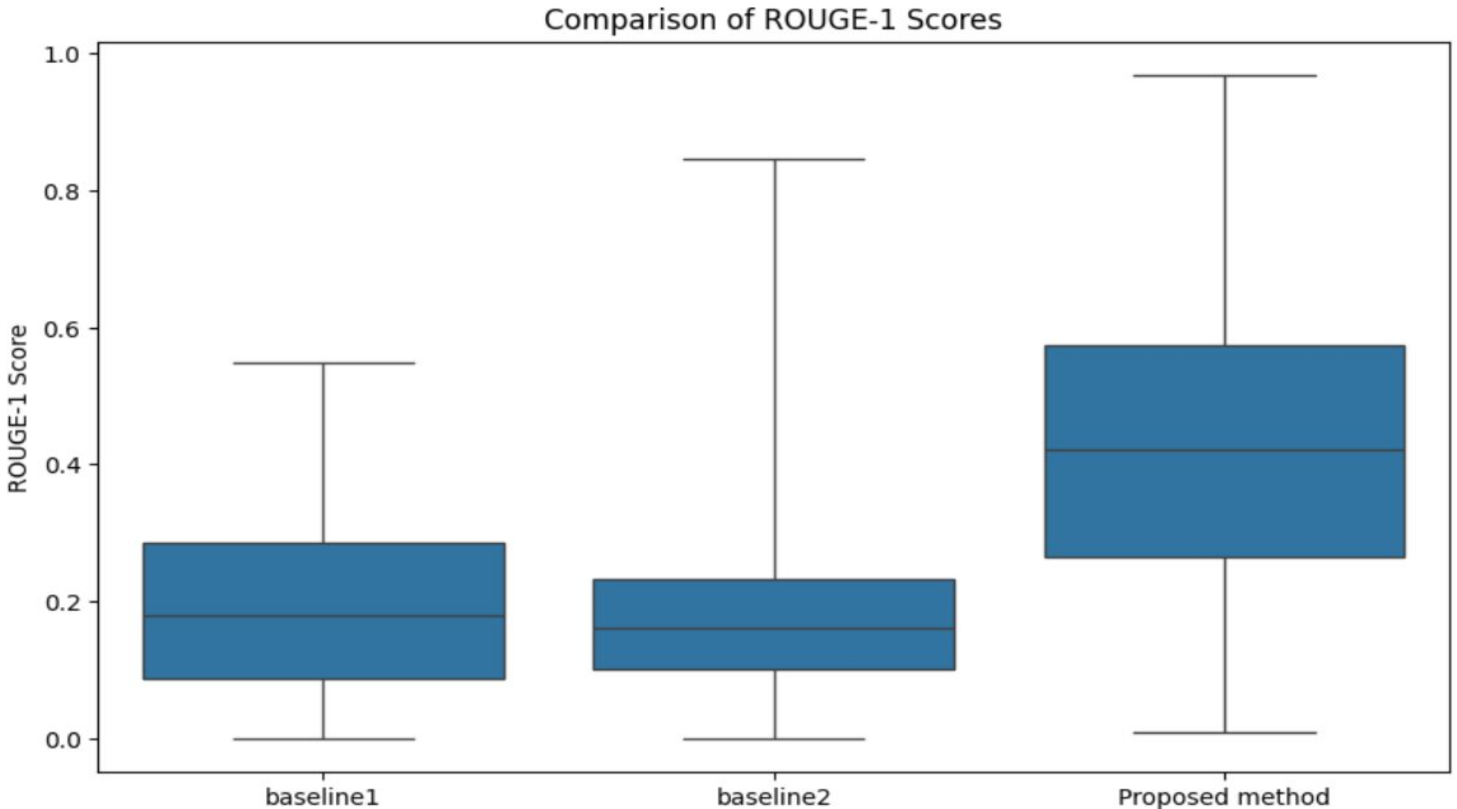
Evaluation Results - Perplexity Comparison of GPT2

- The perplexity decreases which means the output is more coherent.



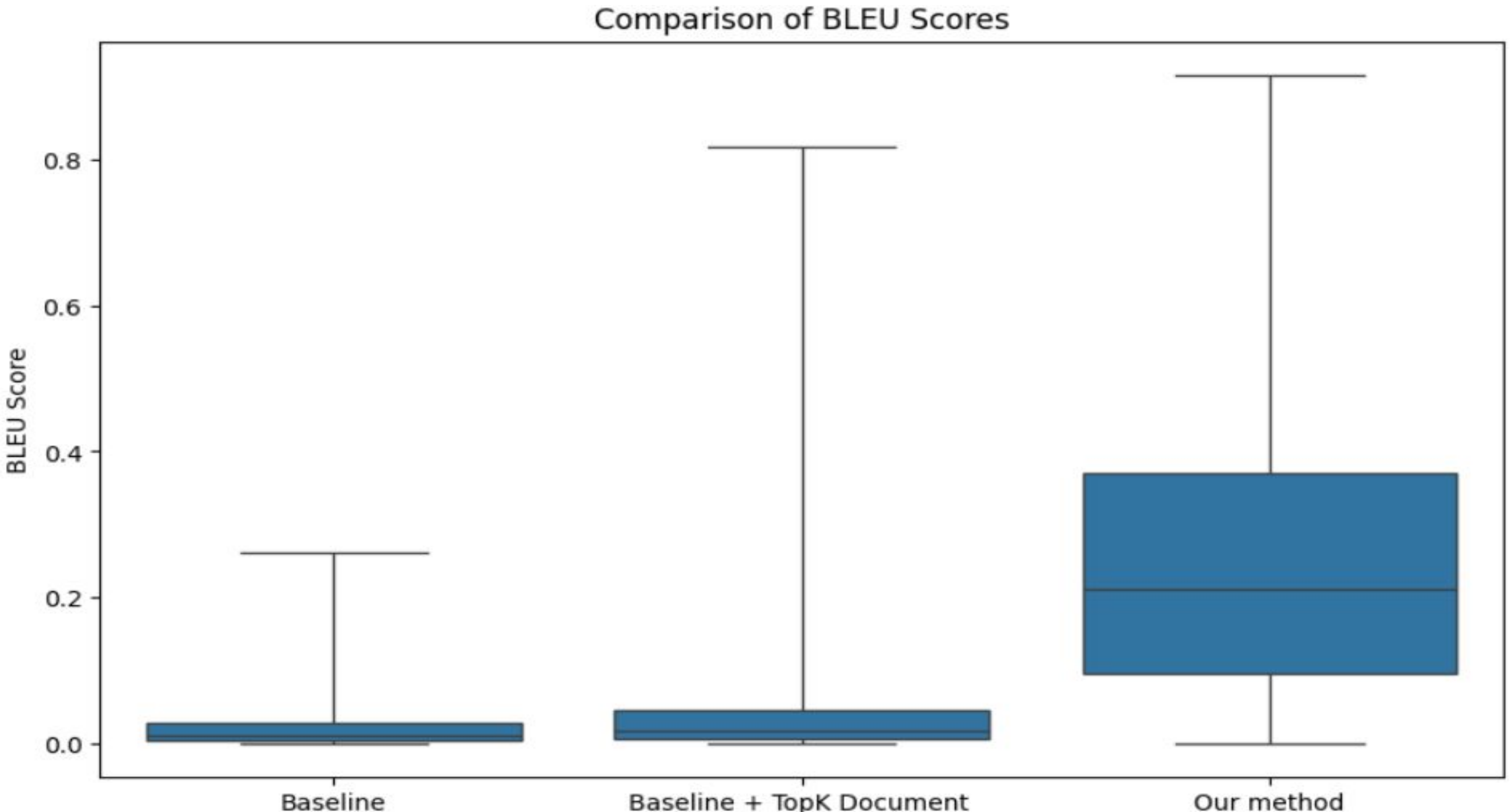
Evaluation Results - ROUGE Comparison on LLaMa2

- LLaMa2 also get higher ROUGE score using relevant sentences.



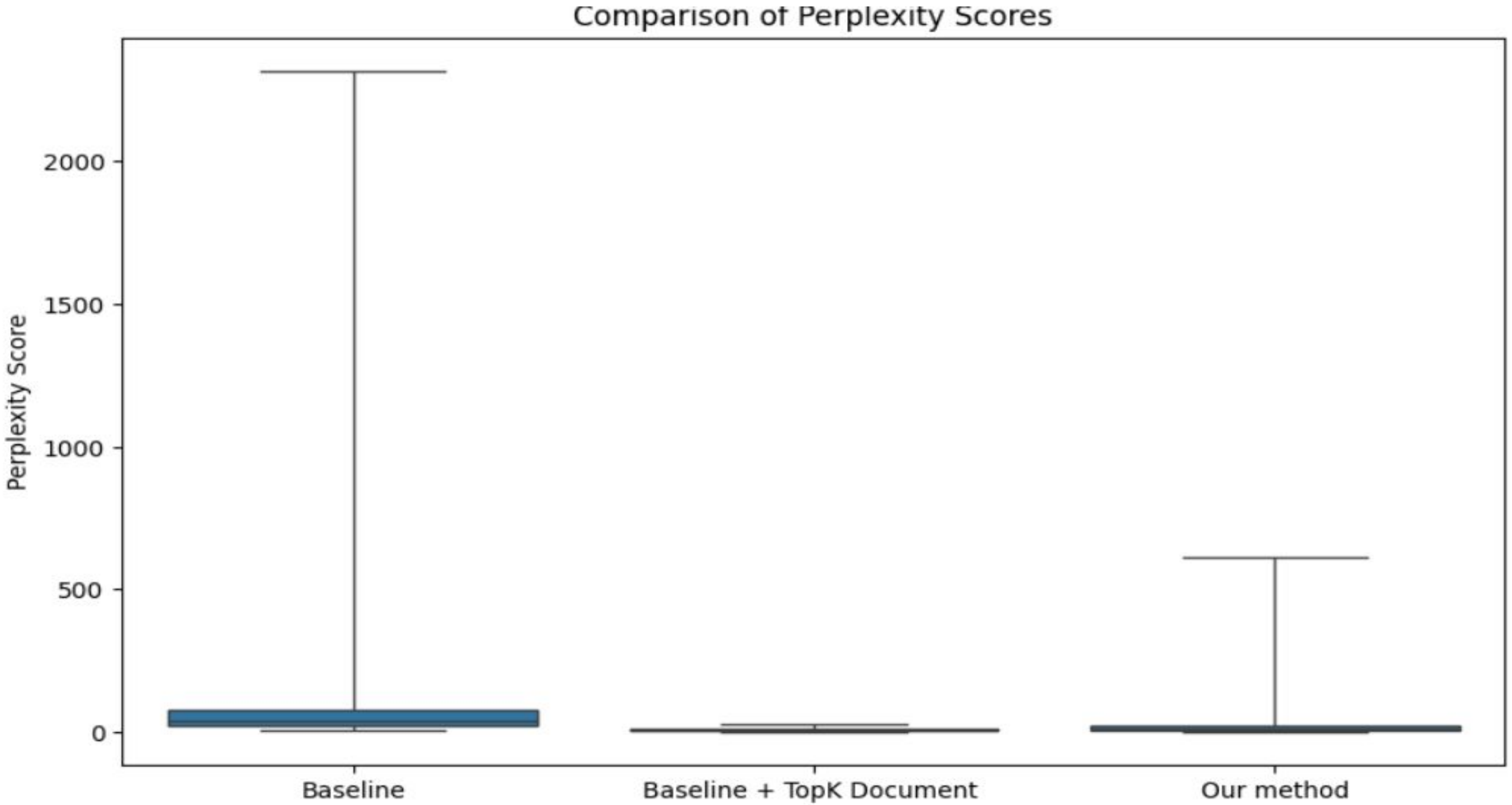
Evaluation Results - BLEU Comparison on LLaMa2

- The BLEU score of LLaMa2 also increases compares with baselines.



Evaluation Results - Perplexity Comparison on LLaMa2

- Topk documents provide more comprehensive context which might boost the coherence.
- Relevant sentences might lead to less coherences since it is a portion of whole context



Difficulties and Limitations

1) ***Long computing time***
    a) No enough computational resources, e.g., GPUs
    b) Large size of wikipedia documents
2) ***Inference problem***
    a) The context size of GPT2 is limited and we can not fit the entire top document into the prompt of GPT2
3) ***Dataset Limitation***
    a) The NQdataset is not comprehensive enough to convey a wide range of knowledge.
4) ***More Evaluation Metrics***
    a) Extra metrics to evaluate the lexical cohesion and word order.

Summary of Contribution

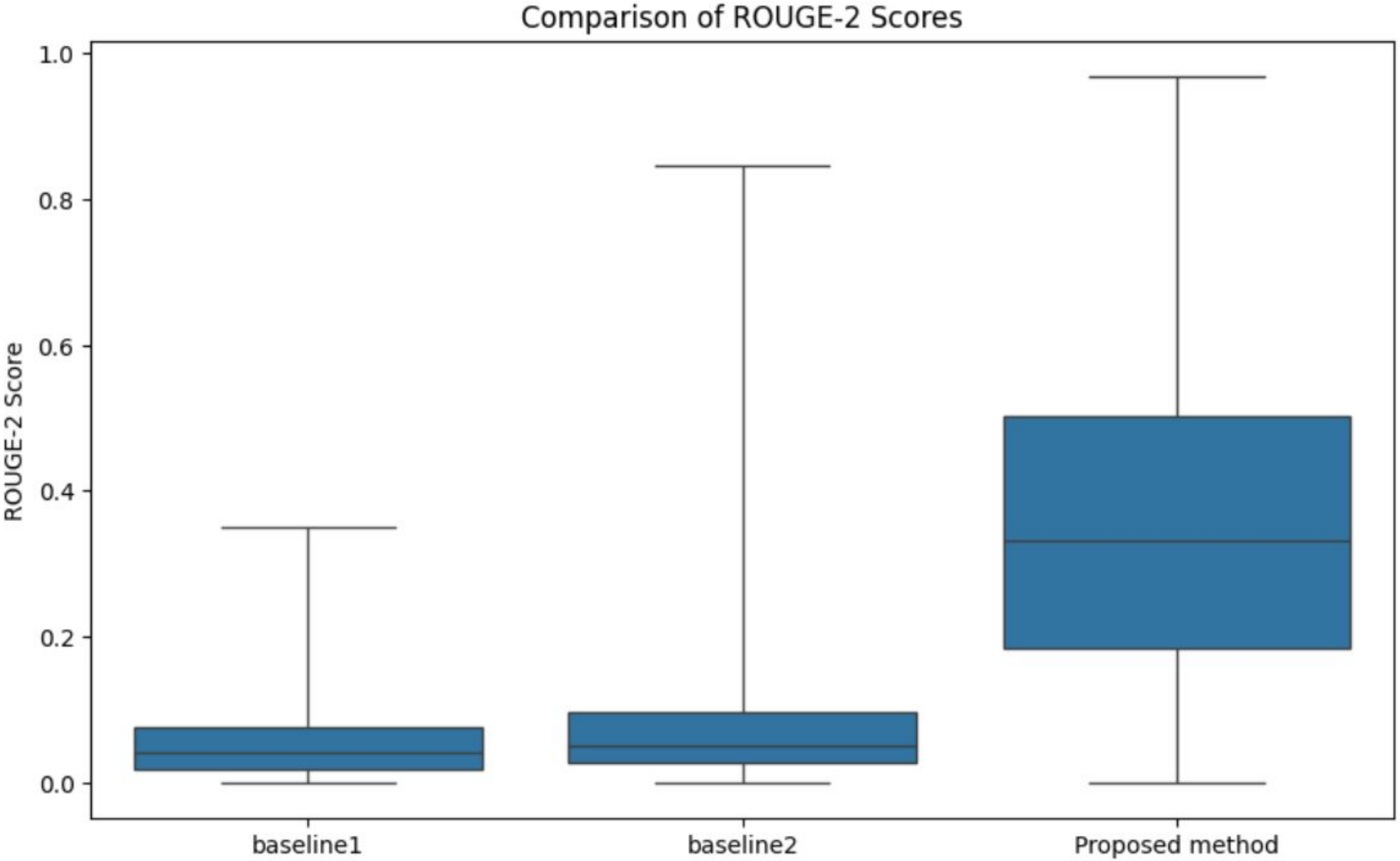
- Proposed a classifier-based method to more accurately extract relevant information from mixed and comprehensive materials for RAG system.
- Conducted extensive evaluation on the proposed method on open-source models including GPT2 and LLaMa2.
- The proposed method achieve better performance in ROUGE, BLUE and Perplexity metrics compared to two baselines.

RAG Sequence and RAG Token

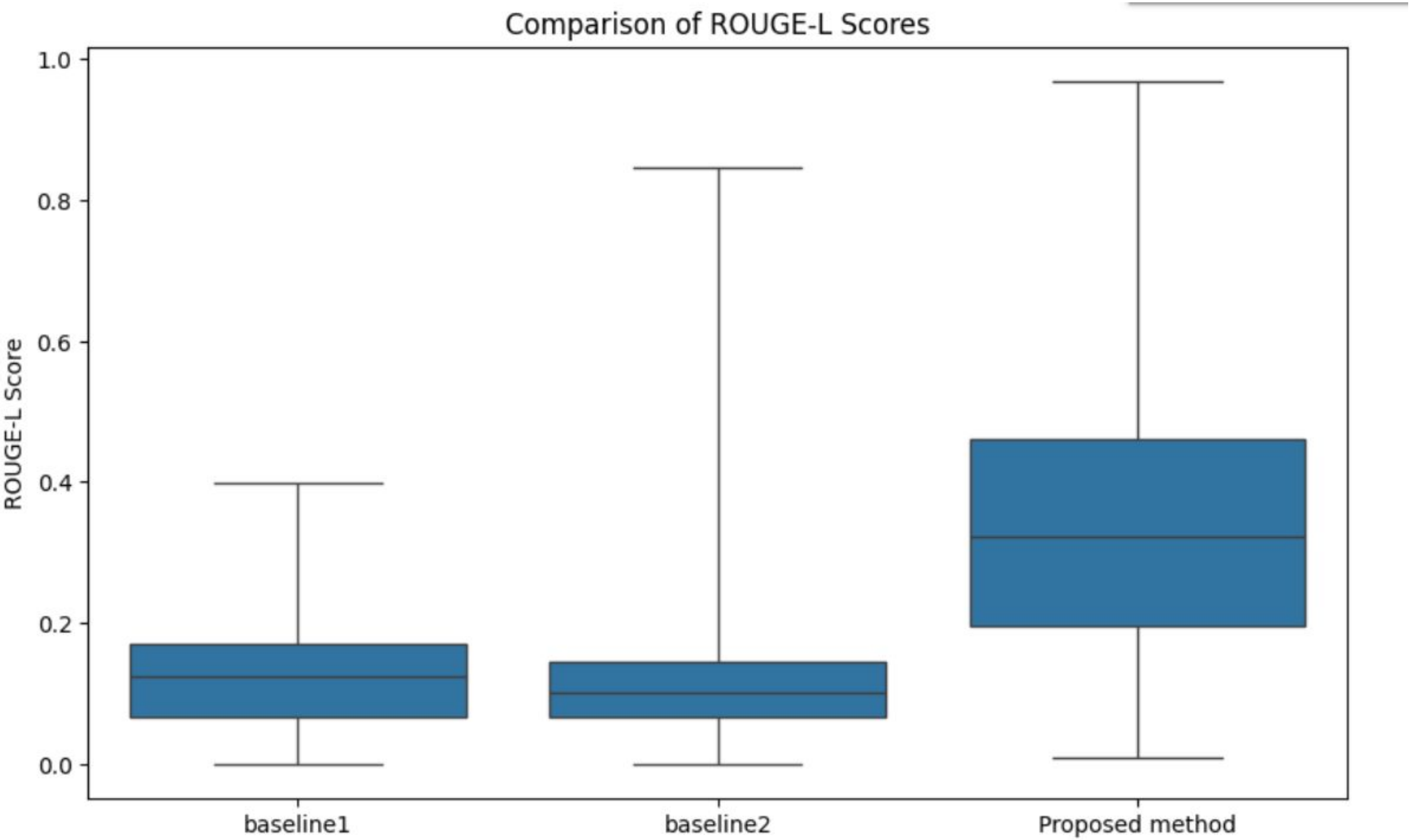
$$p_{\text{RAG-Sequence}}(y|x) \;\approx\; \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) p_\theta(y|x,z) \;=\; \sum_{z \in \text{top-}k(p(\cdot|x))} p_\eta(z|x) \prod_{i}^{N} p_\theta(y_i|x,z,y_{1:i-1})$$

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

Evaluation Results - Rouge2 Comparison on LLaMa2



Evaluation Results - RougeL Comparison on LLaMa2


Comparison of ROUGE-L Scores
ROUGE-L Score
baseline1
baseline2
Proposed method