28.11.23

OPENCAMPUS.sh

Practical Engineering with LLMs

Advanced Retrieval Augmented Generation

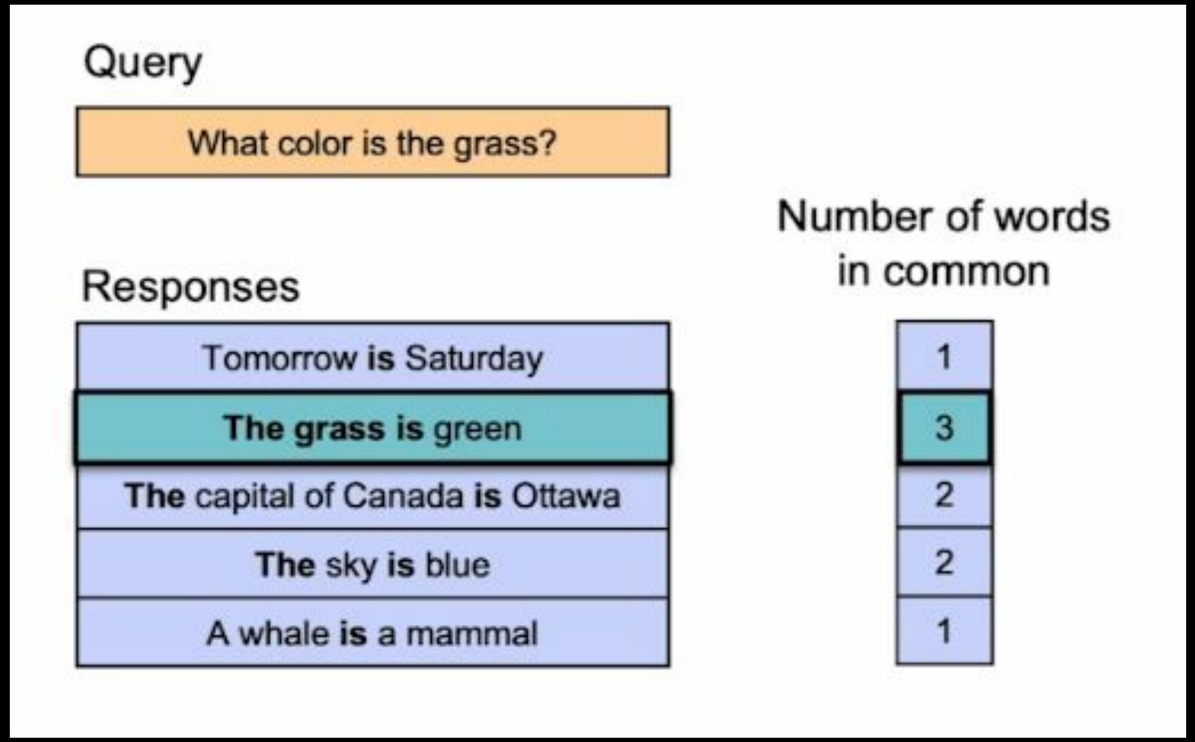
TODAY’S SCHEDULE

- Quiz
- Short Recap
- Project Prototype Presentations
- Homework for next week

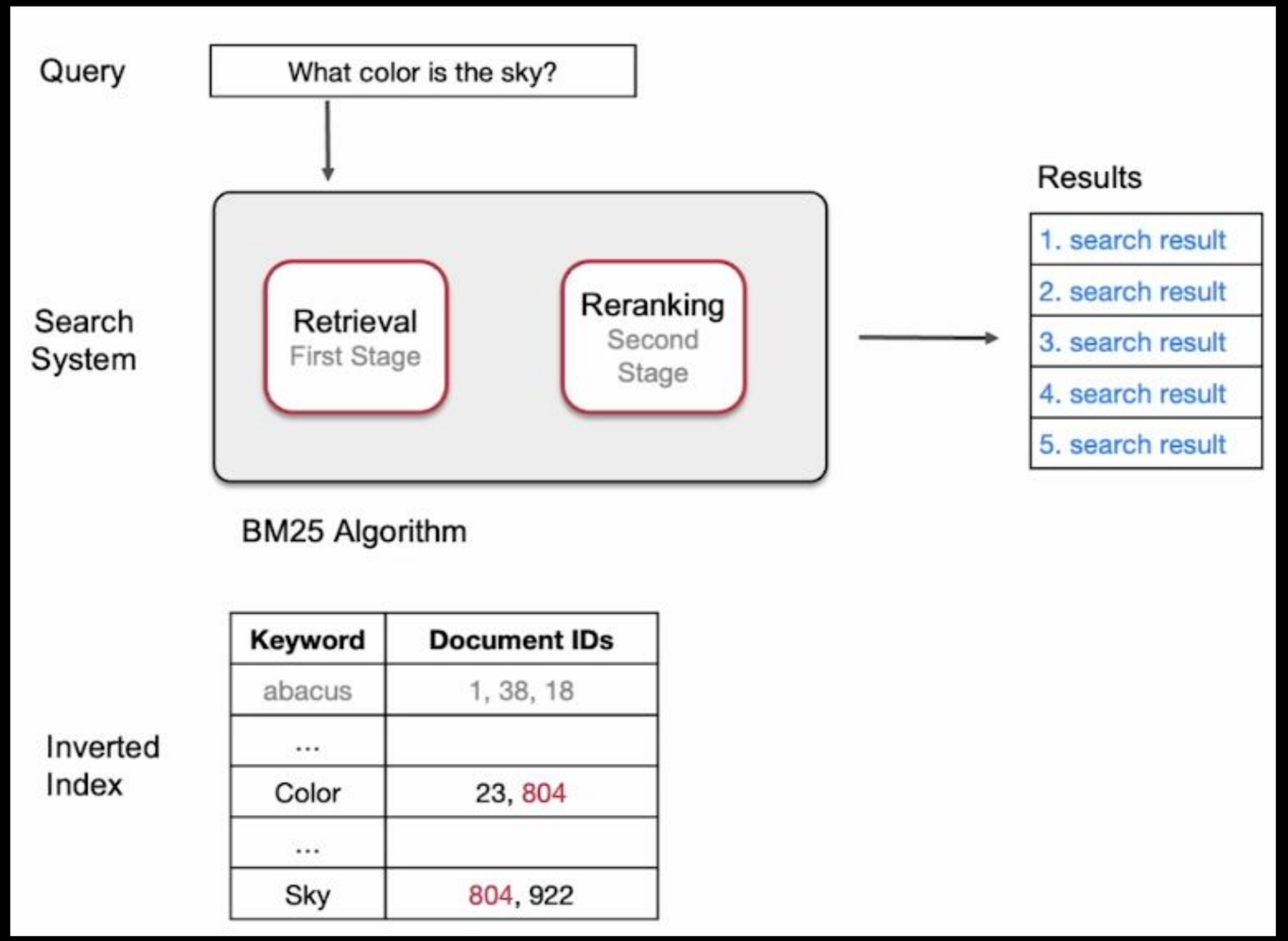
QUIZ

https://forms.office.com/r/LGZcnzQfA9

Keyword Search

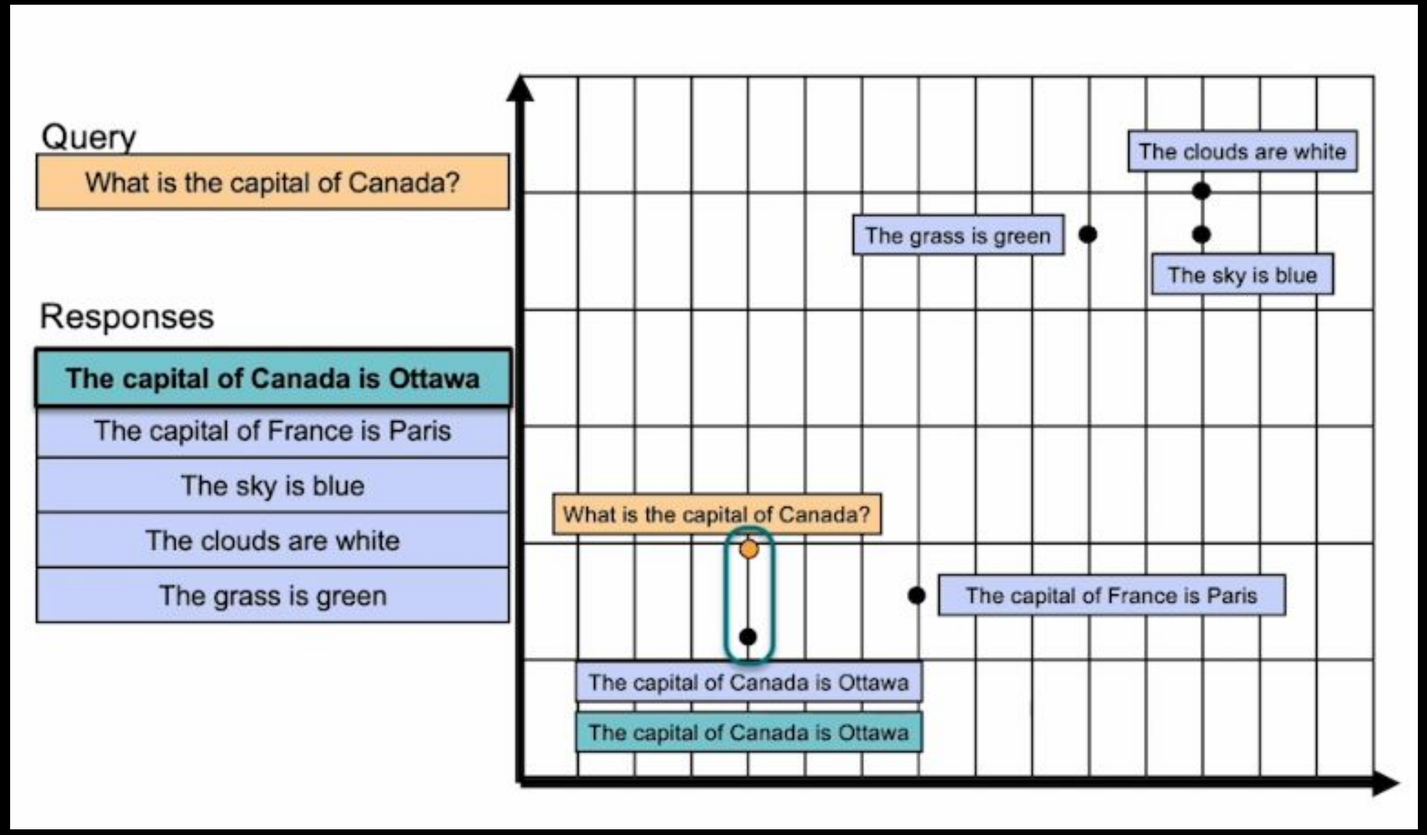


Keyword Search



Vector Search vs. Vector Database

Embeddings


Query
What is the capital of Canada?
Responses
The capital of Canada is Ottawa
The capital of France is Paris
The sky is blue
The clouds are white
The grass is green
The clouds are white
The grass is green
The sky is blue
What is the capital of Canada?
The capital of France is Paris
The capital of Canada is Ottawa
The capital of Canada is Ottawa

Embeddings

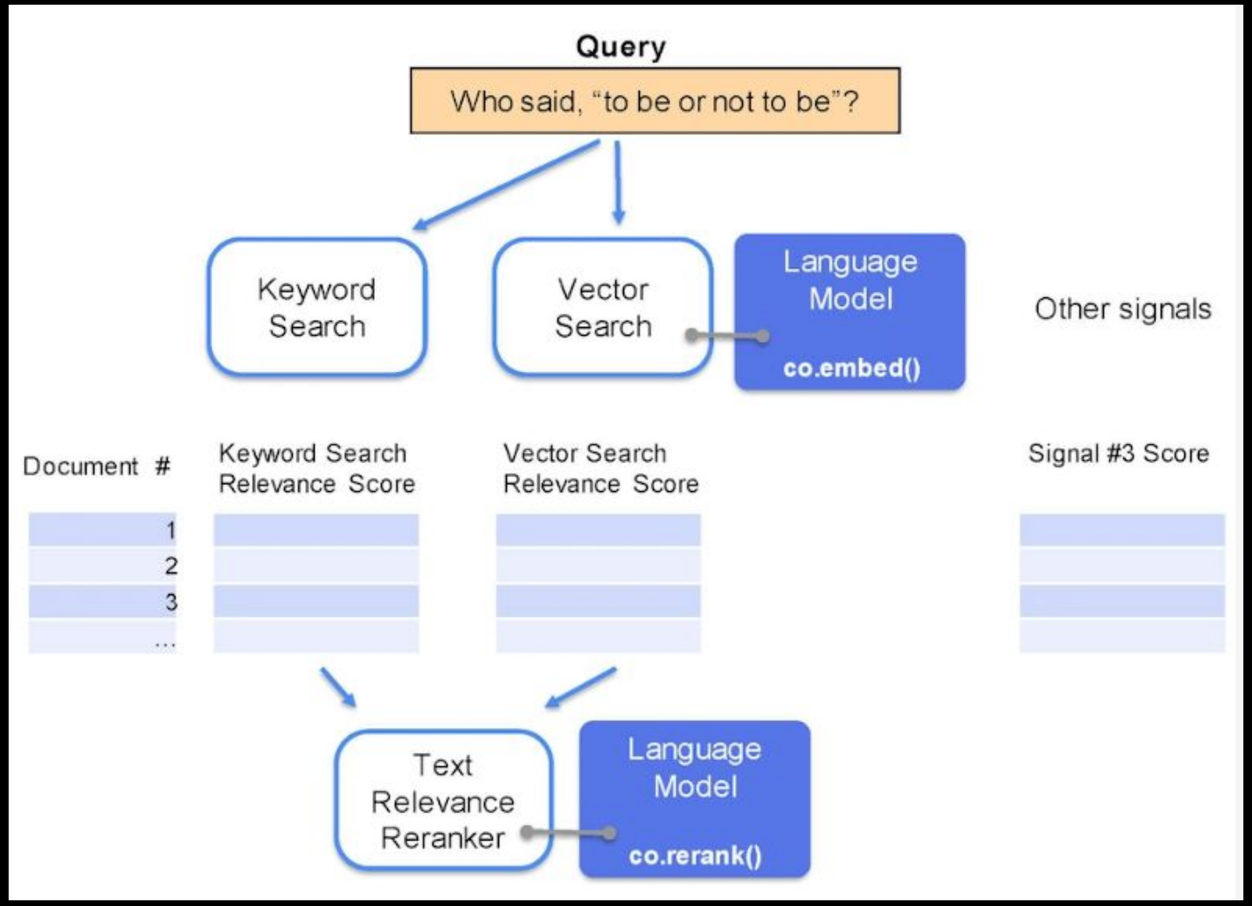
English | Chinese | Polish

Overall MTEB English leaderboard 🔮

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

Rank	Model	Model Size (GB)	Embedding Dimensions	Sequence Length	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)
1	voyage-lite-01-instruct		1024	4096	64.49	74.79	47.4	86.57	59.74	55.58
2	Cohere-embed-english-v3.0		1024	512	64.47	76.49	47.43	85.84	58.01	55
3	bge-large-en-v1.5	1.34	1024	512	64.23	75.97	46.08	87.12	60.03	54.29
4	Cohere-embed-multilingual-v3.0		1024	512	64.01	76.01	46.6	86.15	57.86	53.84

Embeddings


Query
Who said, “to be or not to be”?
Keyword Search
Vector Search
Language Model
co.embed()
Other signals
Document #
Keyword Search Relevance Score
Vector Search Relevance Score
Signal #3 Score
1
2
3
...
Text Relevance Reranker
Language Model
co.rerank()

Embeddings

English | Chinese | Polish

Retrieval English Leaderboard 🔎

- **Metric:** Normalized Discounted Cumulative Gain @ k (ndcg_at_10)
- **Languages:** English

Rank	Model	Average	ArguAna	ArguAna-PL	CQADupstackRetrieval	ClimateFEVER	DBPedia	DBPedia-PL	FEVER
1	multilingual-e5-large	49.73	54.38	53.02	39.68	25.73	41.29	35.82	82.81
2	multilingual-e5-base	46.27	44.23	42.81	38.52	23.86	40.36	30.23	79.44
3	multilingual-e5-small	44.4	39.06	37.43	36.07	22.55	37.76	29.27	75.27
4	paraphrase-multilingual-mpnet-base-v2	31.54	48.91	42.62	31.32	15.27	26.22	20.18	56.76
5	paraphrase-multilingual-MiniLM-L12-v2	28.77	44.88	37.83	30.7	18.49	22.63	18	52.66
6	LaBSE	19.14	34.18	38.52	18.75	3.83	15.57	16.1	12.18
7	text-embedding-ada-002		57.44		41.69	21.64	39.39		74.99
8	text-search-ada-001		46.91			18.5	36.2		72.1
9	text-search-babbage-001		49.2			19.9			77
10	text-search-curie-001		46.98			19.4			75.6

Embeddings

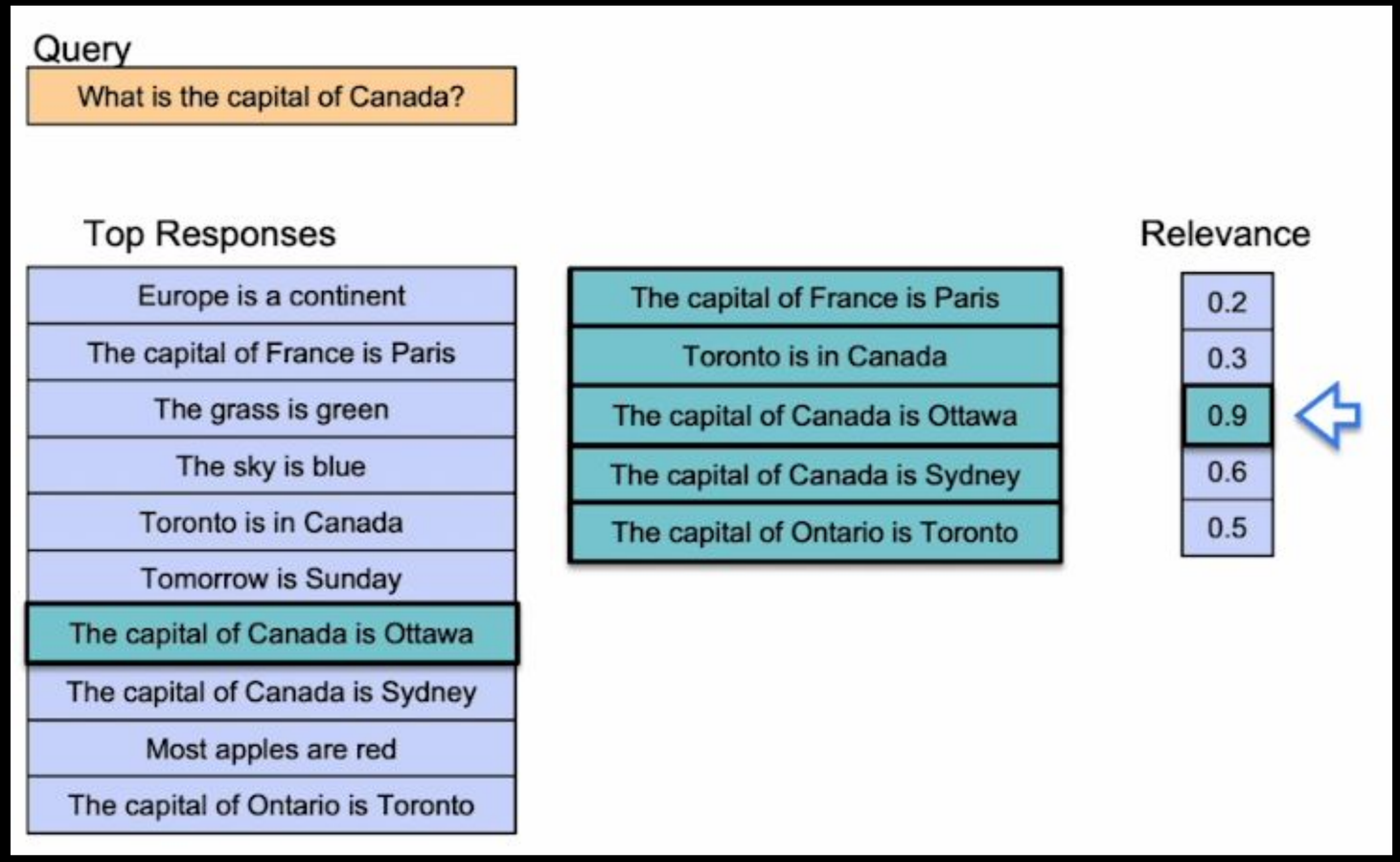
English | Chinese | Polish

Retrieval English Leaderboard 🔎

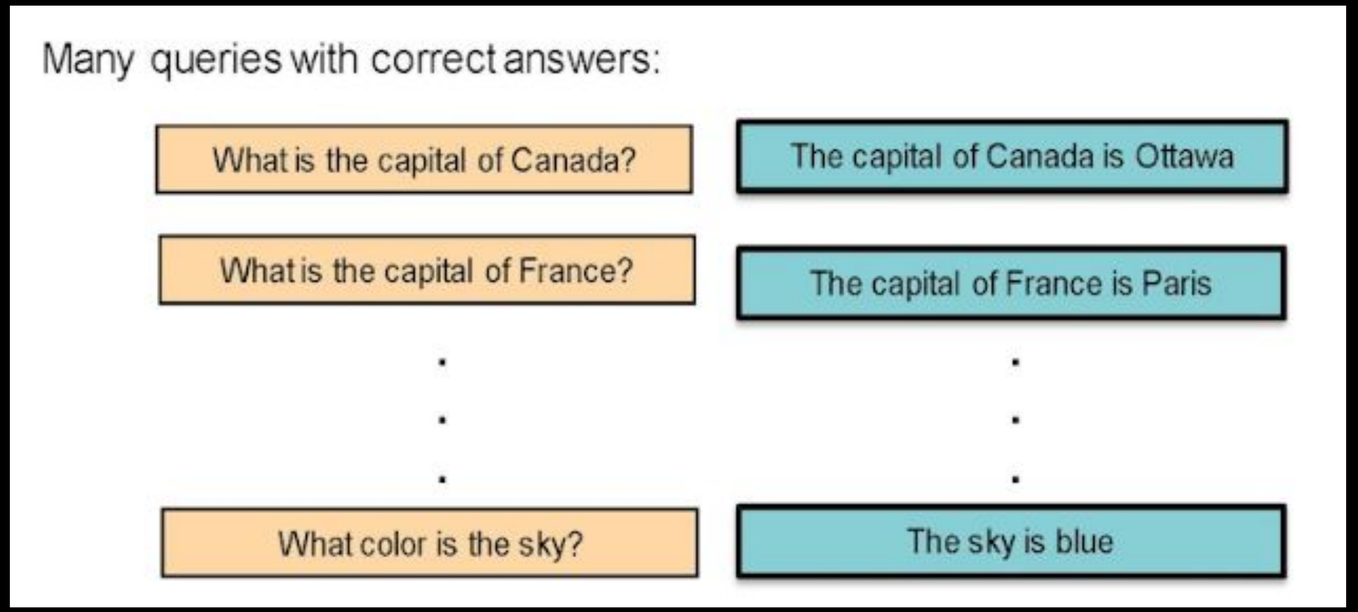
- **Metric:** Normalized Discounted Cumulative Gain @ k (ndcg_at_10)
- **Languages:** English

Rank	Model	Average	ArguAna	ArguAna-PL	CQADupstackRetrieval	ClimateFEVER	DBPedia	DBPedia-PL	FEVER
67	ember-v1		64.56		42.39	27.29	41.79		83.69
58	sf_model_e5		64.07		41.14	28.74	42.51		81.38
17	bge-base-en-v1.5		63.61		42.35	31.17	40.77		86.29
36	bge-base-en-v1-5-seqlen-384-bs-1		63.61			31.17	40.77		86.29
18	bge-large-en-v1.5		63.54		42.23	36.57	44.11		87.18
21	Cohere-embed-english-v3.0		61.52		41.53	38.43	43.36		88.97
46	stella-base-en-v2		60.63		41.14	29	39.64		79.13
19	bge-small-en-v1.5		59.55		39.05	31.84	40.03		86.64
15	voyage-lite-01-instruct		58.73		45.11	37.47	43.42		89.71
68	bge-small-en-v1.5-quant		57.77						

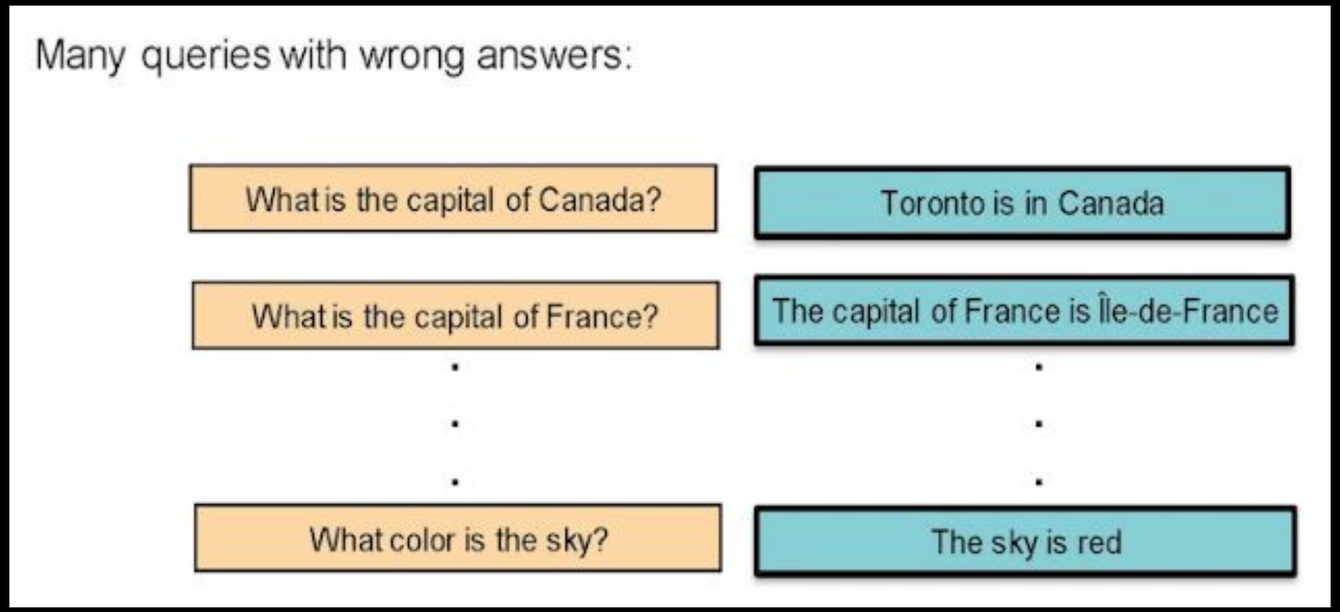
Reranking


Query
What is the capital of Canada?
Top Responses
Europe is a continent
The capital of France is Paris
The grass is green
The sky is blue
Toronto is in Canada
Tomorrow is Sunday
The capital of Canada is Ottawa
The capital of Canada is Sydney
Most apples are red
The capital of Ontario is Toronto
The capital of France is Paris
Toronto is in Canada
The capital of Canada is Ottawa
The capital of Canada is Sydney
The capital of Ontario is Toronto
Relevance
0.2
0.3
0.9
0.6
0.5

Reranking


Many queries with correct answers:
What is the capital of Canada?
The capital of Canada is Ottawa
What is the capital of France?
The capital of France is Paris
What color is the sky?
The sky is blue

Reranking


Many queries with wrong answers:
What is the capital of Canada?
Toronto is in Canada
What is the capital of France?
The capital of France is Île-de-France
What color is the sky?
The sky is red

Reranking

English | Chinese

Reranking English Leaderboard

- **Metric:** Mean Average Precision (MAP)
- **Languages:** English

Rank	Model	Average	AskUbuntuDupQuestions	MindSmallReranking	SciDocsRR	StackOverflowDupQuestions
1	ember-v1	60.04	64.46	32.27	87.56	55.85
2	bge-large-en-v1.5	60.03	64.47	32.06	87.63	55.95
3	sf_model_e5	59.86	64.32	32.27	87.47	55.4
4	voyage-lite-01-instruct	59.74	65.77	31.69	87.03	54.49
5	all-mpnet-base-v2	59.36	65.85	30.97	88.65	51.98
6	gte-large	59.13	63.06	32.63	87.2	53.63
7	bge-base-en-v1.5-quant	58.94	62.39	31.89	87.05	54.45
8	bge-base-en-v1-5-seqlen-384-bs-1	58.86	62.13	31.2	87.49	54.61
9	bge-base-en-v1.5	58.86	62.13	31.2	87.49	54.61
10	stella-base-en-v2	58.78	62.72	31.91	86.66	53.81

Evaluation of retrieval/ranking systems

Mean Average Precision (MAP)

Mean Reciprocal Rank (MRR)

Normalized Discounted Cumulative Gain (NDCG)

$$\mathrm{mAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i$$

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}.$$

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

$$\mathrm{IDCG_p} = \sum_{i=1}^{|REL_p|} \frac{rel_i}{\log_2(i+1)}$$

HyDE (Hypothetical Document Embeddings)

Problems with Embeddings:

- **Query and document content don’t have the same structure**
- **Query may not contain all the necessary information to create a good embedding**

HyDE Solution:

1. **Let the LLM answer the question first**
2. **Embed the generated answer and retrieve the documents**
3. **The answer should contain the key ideas of the answer but doesn’t have to be 100% accurate**

HyDE (Hypothetical Document Embeddings)

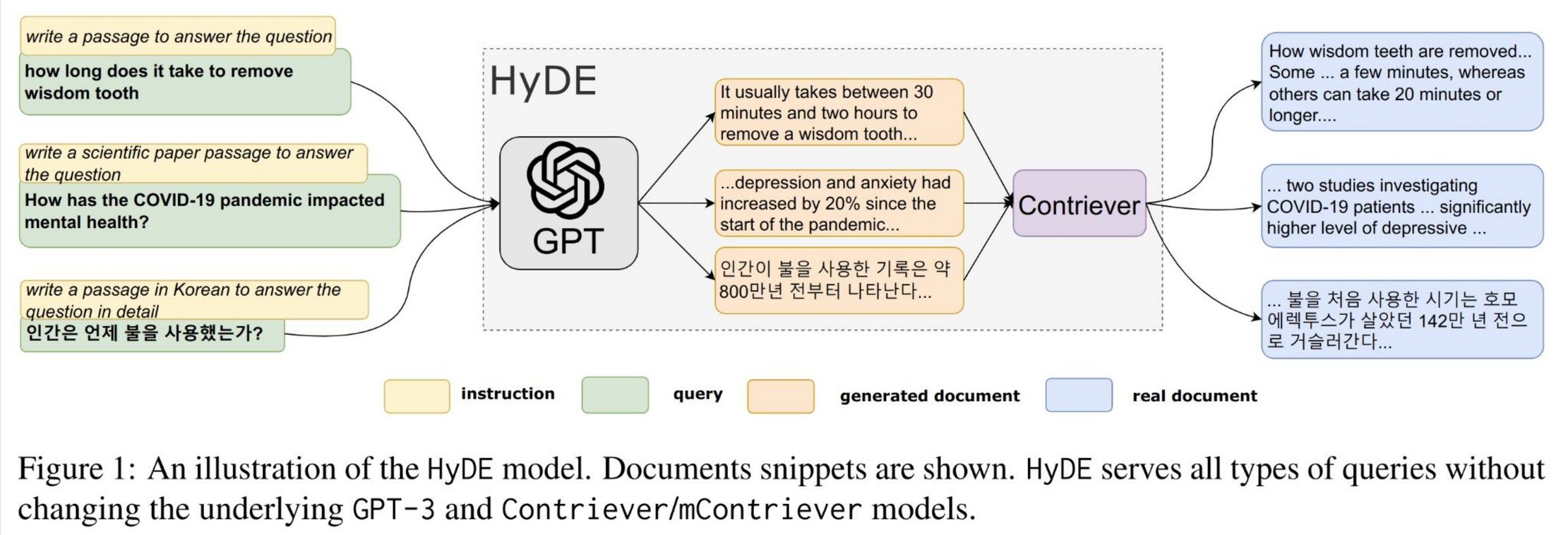


Figure 1: An illustration of the HyDE model. Documents snippets are shown. HyDE serves all types of queries without changing the underlying GPT-3 and Contriever/mContriever models.

https://github.com/texttron/hyde

Sparse Priming Representations (SPR)

Sparse Priming Representations (SPR)

If your native language is English (or not German) look here!

Wenn Deine Muttersprache Deutsch ist, schau hier hin!

Sparse Priming Representations (SPR)

Which of the following words describes a place?

1. Banana
2. Pineapple
3. Zoo
4. Mango

Welches der folgenden Wörter beschreibt einen Ort?

1. Banane
2. Ananas
3. Zoo
4. Mango

Sparse Priming Representations (SPR)

Please write down 10 animals you can think of!

Bitte schreib 10 Tiere auf, die Dir einfallen!

Sparse Priming Representations (SPR)

Are these in your list?

Lion
Elephant
Dolphin
Penguin
Giraffe
Koala
Tiger
Wolf
Kangaroo
Cheetah

Sind diese in Deiner Liste?

Löwe
Elefant
Delfin
Pinguin
Giraffe
Koala
Tiger
Wolf
Känguru
Gepard

Sparse Priming Representations (SPR)

Which of the following words describes a place?

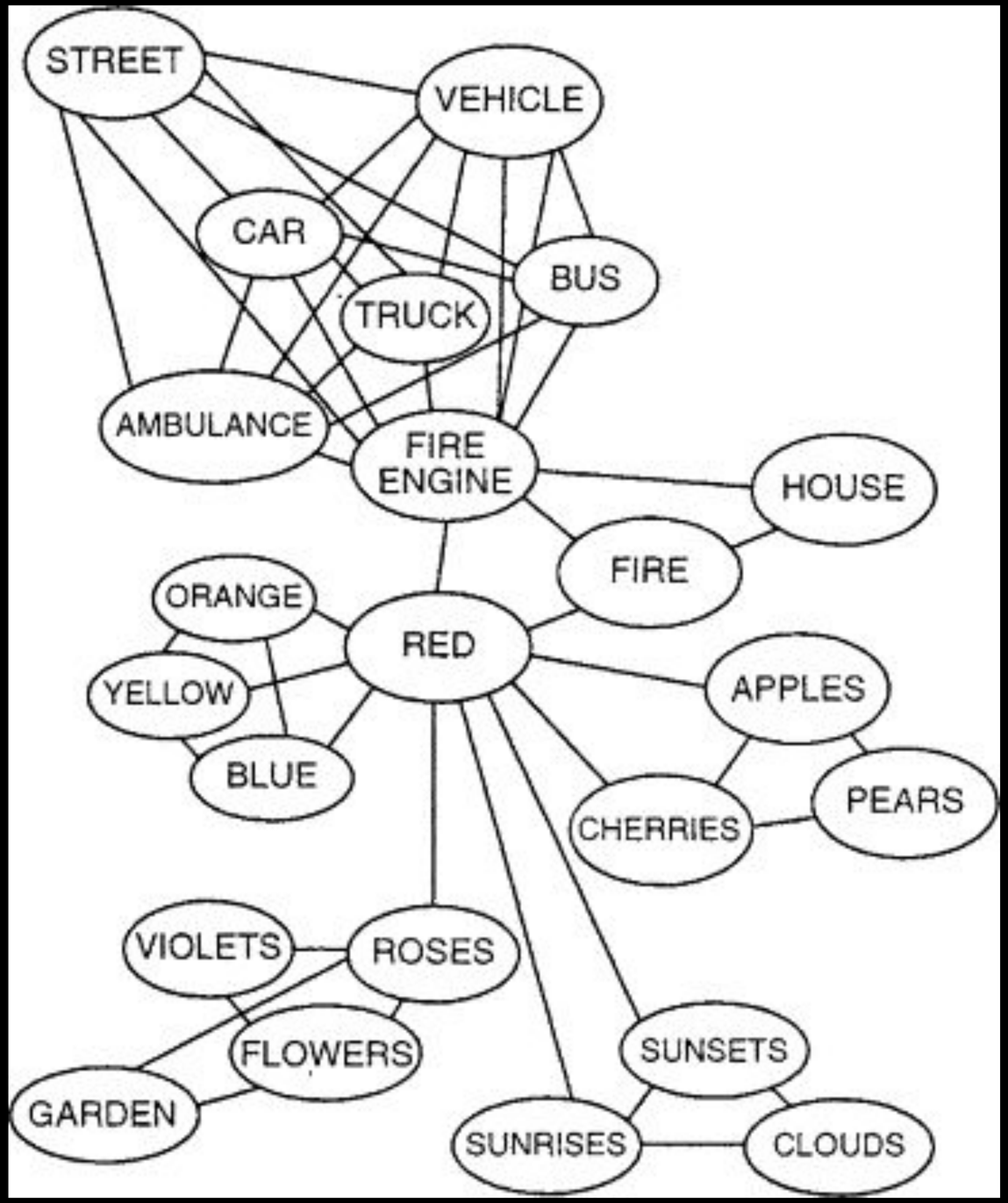
1. Pet
2. Bone
3. Garage
4. Leash

Sparse Priming Representations (SPR)

Are these in your list?

Dog
Cat
Elephant
Lion
Dolphin
Giraffe
Penguin
Tiger
Bear
Kangaroo

Sparse Priming Representations (SPR)



Sparse Priming Representations (SPR)

SPR Generator

MISSION
You are a Sparse Priming Representation (SPR) writer. An SPR is a particular kind of use of language for advanced NLP, NLU, and NLG tasks, particularly useful for the latest generation of Large Language Models (LLMs). You will be given information by the USER which you are to render as an SPR.

THEORY
LLMs are a kind of deep neural network. They have been demonstrated to embed knowledge, abilities, and concepts, ranging from reasoning to planning, and even to theory of mind. These are called latent abilities and latent content, collectively referred to as latent space. **The latent space of an LLM can be activated with the correct series of words as inputs**, which will create a useful internal state of the neural network. This is not unlike how the right shorthand cues can prime a human mind to think in a certain way. Like human minds, LLMs are associative, meaning you only need to use the correct associations to "prime" another model to think in the same way.

METHODOLOGY
Render the input as a distilled list of succinct statements, assertions, associations, concepts, analogies, and metaphors. The idea is to capture as much, conceptually, as possible but with as few words as possible. Write it in a way that makes sense to you, as the future audience will be another language model, not a human. Use complete sentences.

Sparse Priming Representations (SPR)

SPR Decompressor

MISSION
You are a Sparse Priming Representation (SPR) decompressor. An SPR is a particular kind of use of language for advanced NLP, NLU, and NLG tasks, particularly useful for the latest generation of Large Language Models (LLMs). You will be given an SPR and your job is to fully unpack it.

THEORY
LLMs are a kind of deep neural network. They have been demonstrated to embed knowledge, abilities, and concepts, ranging from reasoning to planning, and even to theory of mind. These are called latent abilities and latent content, collectively referred to as latent space. **The latent space of an LLM can be activated with the correct series of words as inputs,** which will create a useful internal state of the neural network. This is not unlike how the right shorthand cues can prime a human mind to think in a certain way. Like human minds, LLMs are associative, meaning you only need to use the correct associations to "prime" another model to think in the same way.

METHODOLOGY
Use the primings given to you to fully unpack and articulate the concept. Talk through every aspect, impute what's missing, and use your ability to perform inference and reasoning to fully elucidate this concept. Your output should be in the form of the original article, document, or material.

Sparse Priming Representations (SPR)

The lion (Panthera leo) is a large cat of the genus Panthera native to Africa and India. It has a muscular, broad-chested body; short, rounded head; round ears; and a hairy tuft at the end of its tail. It is sexually dimorphic; adult male lions are larger than females and have a prominent mane. It is a social species, forming groups called prides. A lion's pride consists of a few adult males, related females, and cubs. Groups of female lions usually hunt together, preying mostly on large ungulates. The lion is an apex and keystone predator; although some lions scavenge when opportunities occur and have been known to hunt humans, lions typically do not actively seek out and prey on humans.

Sparse Priming Representations (SPR)

Panthera leo: large cat, Africa, India.
Physical traits: muscular, broad-chested, short-rounded head, round ears, hairy tail tuft.
Sexual dimorphism: males larger, mane.
Social structure: prides - adult males, related females, cubs.
Hunting: female groups, large ungulates, apex and keystone predator, occasional human prey.
Habitat: grasslands, savannahs, shrublands.
Activity: diurnal, adapts to nocturnal if threatened.
Historical range: Africa, Eurasia, Southeast Europe to India.
Current status: fragmented, sub-Saharan Africa, western India.
Conservation: Vulnerable, 43% decline since 1990s

Sparse Priming Representations (SPR)

Panthera leo, commonly known as the lion, is a prominent large cat species found in the diverse landscapes of Africa and India. Characterized by their robust, muscular build and broad chests, lions possess a distinct physical appearance with their short, rounded heads, round ears, and a distinctive tuft at the end of their tails. A key aspect of their identity is sexual dimorphism; males are notably larger and boast impressive manes, setting them apart from their female counterparts. Lions have a unique social structure, living in groups called prides, consisting of adult males and females, as well as their cubs. Hunting is a pivotal part of their lifestyle, predominantly a group activity led by females, targeting large ungulates like zebras and wildebeests. As apex predators and keystone species in their ecosystems, lions play a critical ecological role, maintaining the balance by regulating prey populations. They are occasional scavengers and, in rare instances, may hunt humans. Their status as top predators and their significant impact on their habitats underscore their vital role in the ecological framework.

Sparse Priming Representations (SPR)

The lion (Panthera leo) is a large cat of the genus Panthera native to Africa and India. It has a muscular, broad-chested body; short, rounded head; round ears; and a hairy tuft at the end of its tail. It is sexually dimorphic; adult male lions are larger than females and have a prominent mane. It is a social species, forming groups called prides. A lion's pride consists of a few adult males, related females, and cubs. Groups of female lions usually hunt together, preying mostly on large ungulates. The lion is an apex and keystone predator; although some lions scavenge when opportunities occur and have been known to hunt humans, lions typically do not actively seek out and prey on humans.

Panthera leo, commonly known as the lion, is a prominent large cat species found in the diverse landscapes of Africa and India. Characterized by their robust, muscular build and broad chests, lions possess a distinct physical appearance with their short, rounded heads, round ears, and a distinctive tuft at the end of their tails. A key aspect of their identity is sexual dimorphism; males are notably larger and boast impressive manes, setting them apart from their female counterparts. Lions have a unique social structure, living in groups called prides, consisting of adult males and females, as well as their cubs. Hunting is a pivotal part of their lifestyle, predominantly a group activity led by females, targeting large ungulates like zebras and wildebeests. As apex predators and keystone species in their ecosystems, lions play a critical ecological role, maintaining the balance by regulating prey populations. They are occasional scavengers and, in rare instances, may hunt humans. Their status as top predators and their significant impact on their habitats underscore their vital role in the ecological framework.

Project Presentations

One person per group presents

1. the provided project template
2. if available a small first prototype of the project idea

Homework Until Next Week

- Watch the course “Building Generative AI Applications with Gradio”
- Install Gradio and explore the examples from the Gradio playground locally (https://www.gradio.app/playground)
- Think about the elements your application user interface should contain, create a sketch or wireframe of your app and implement it with Gradio
- (If possible connect your prototype to the user interface)