Information Retrieval

Oct 1, 2024 @ Introduction to Human Language Technology

Eugene Yang eugene.yang@jhu.edu

Slides borrowed from SIGIR24 Tutorial
“Neural Methods for Cross-Language Information Retrieval”

What is Information Retrieval?

(relevant)

Retrieve information from a storage based on user's information need

Don’t we have Google?

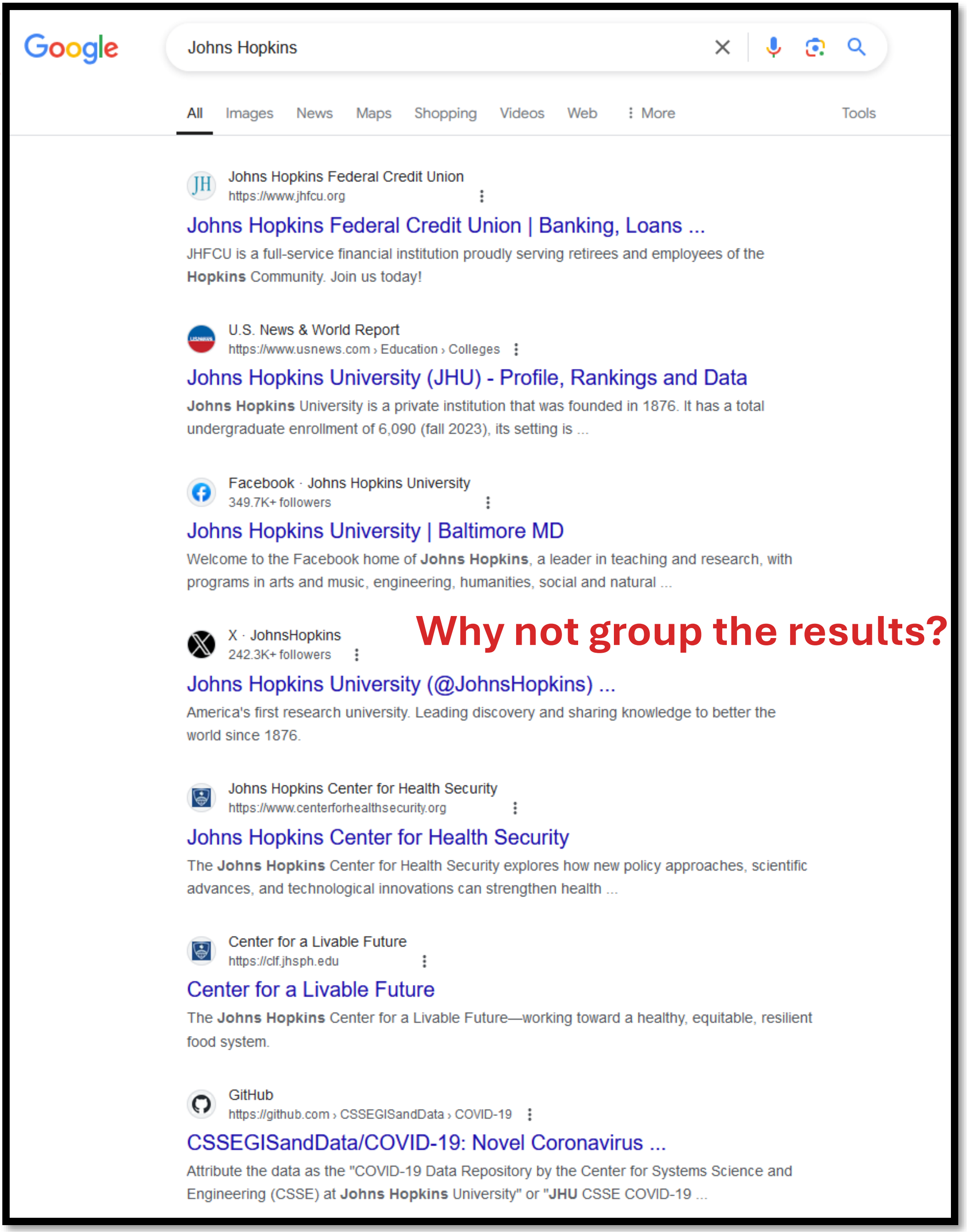
Yes, **but Google is not all.**

What if I’m looking for the person?

Why not read my mind?

Why in a list?

Why asking me to read?



Johns Hopkins Federal Credit Union
https://www.jhfcu.org

Johns Hopkins Federal Credit Union | Banking, Loans ...

JHFCU is a full-service financial institution proudly serving retirees and employees of the **Hopkins** Community. Join us today!

U.S. News & World Report
https://www.usnews.com › Education › Colleges

Johns Hopkins University (JHU) - Profile, Rankings and Data

Johns Hopkins University is a private institution that was founded in 1876. It has a total undergraduate enrollment of 6,090 (fall 2023), its setting is ...

Facebook · Johns Hopkins University
349.7K+ followers

Johns Hopkins University | Baltimore MD

Welcome to the Facebook home of **Johns Hopkins**, a leader in teaching and research, with programs in arts and music, engineering, humanities, social and natural ...

X · JohnsHopkins
242.3K+ followers

Johns Hopkins University (@JohnsHopkins) ...

America's first research university. Leading discovery and sharing knowledge to better the world since 1876.

Johns Hopkins Center for Health Security
https://www.centerforhealthsecurity.org

Johns Hopkins Center for Health Security

The **Johns Hopkins** Center for Health Security explores how new policy approaches, scientific advances, and technological innovations can strengthen health ...

Center for a Livable Future
https://clf.jhsph.edu

Center for a Livable Future

The **Johns Hopkins** Center for a Livable Future—working toward a healthy, equitable, resilient food system.

GitHub
https://github.com › CSSEGISandData › COVID-19

CSSEGISandData/COVID-19: Novel Coronavirus ...

Attribute the data as the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at **Johns Hopkins** University" or "**JHU** CSSE COVID-19 ...

Why not group the results?

Google Search is just one implementation

Google trained us well!

- Even faster?
- Smarter?
- Cross language?

Hard Matching Problem

- Text to text
  - Search in notes
  - Cross language search
  - Cross domain search
- Text to other modalities
  - Image search
  - Video search

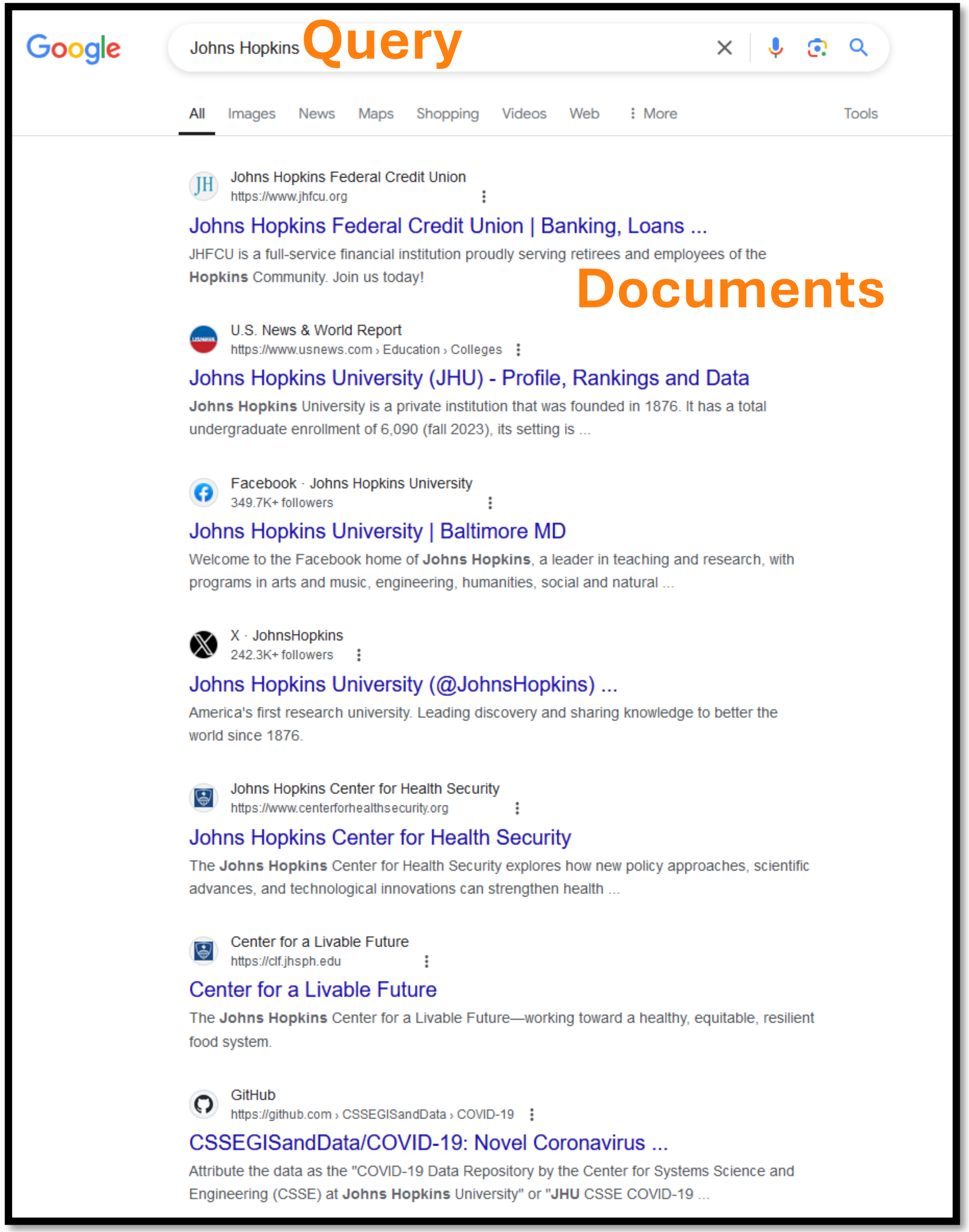
Different Search Process

- Iterative search
  - e.g., electronic discovery and systematic review
- Conversational search
  - Alexa search
- Recommendation systems
  - Implicit queries
- (Set Retrieval)

Core Problem

- Rank relevant document at top
- Do it fast

Ranked List



Query

Documents

Johns Hopkins Federal Credit Union
https://www.jhfcu.org

Johns Hopkins Federal Credit Union | Banking, Loans ...

JHFCU is a full-service financial institution proudly serving retirees and employees of the **Hopkins** Community. Join us today!

U.S. News & World Report
https://www.usnews.com › Education › Colleges

Johns Hopkins University (JHU) - Profile, Rankings and Data

Johns Hopkins University is a private institution that was founded in 1876. It has a total undergraduate enrollment of 6,090 (fall 2023), its setting is ...

Facebook · Johns Hopkins University
349.7K+ followers

Johns Hopkins University | Baltimore MD

Welcome to the Facebook home of **Johns Hopkins**, a leader in teaching and research, with programs in arts and music, engineering, humanities, social and natural ...

X · JohnsHopkins
242.3K+ followers

Johns Hopkins University (@JohnsHopkins) ...

America's first research university. Leading discovery and sharing knowledge to better the world since 1876.

Johns Hopkins Center for Health Security
https://www.centerforhealthsecurity.org

Johns Hopkins Center for Health Security

The **Johns Hopkins** Center for Health Security explores how new policy approaches, scientific advances, and technological innovations can strengthen health ...

Center for a Livable Future
https://clf.jhsph.edu

Center for a Livable Future

The **Johns Hopkins** Center for a Livable Future—working toward a healthy, equitable, resilient food system.

GitHub
https://github.com › CSSEGISandData › COVID-19

CSSEGISandData/COVID-19: Novel Coronavirus ...

Attribute the data as the "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at **Johns Hopkins** University" or "**JHU** CSSE COVID-19 ...

Design Space

Effectiveness

- Definition of relevancy
- How to model relevancy

Efficiency

- How fast
- Fast at what stage

Agenda

- What is information retrieval?
- Retrieval Modeling and Pipeline
  - Statistical and Neural
- Evaluation
- State of IR Research and active research problems

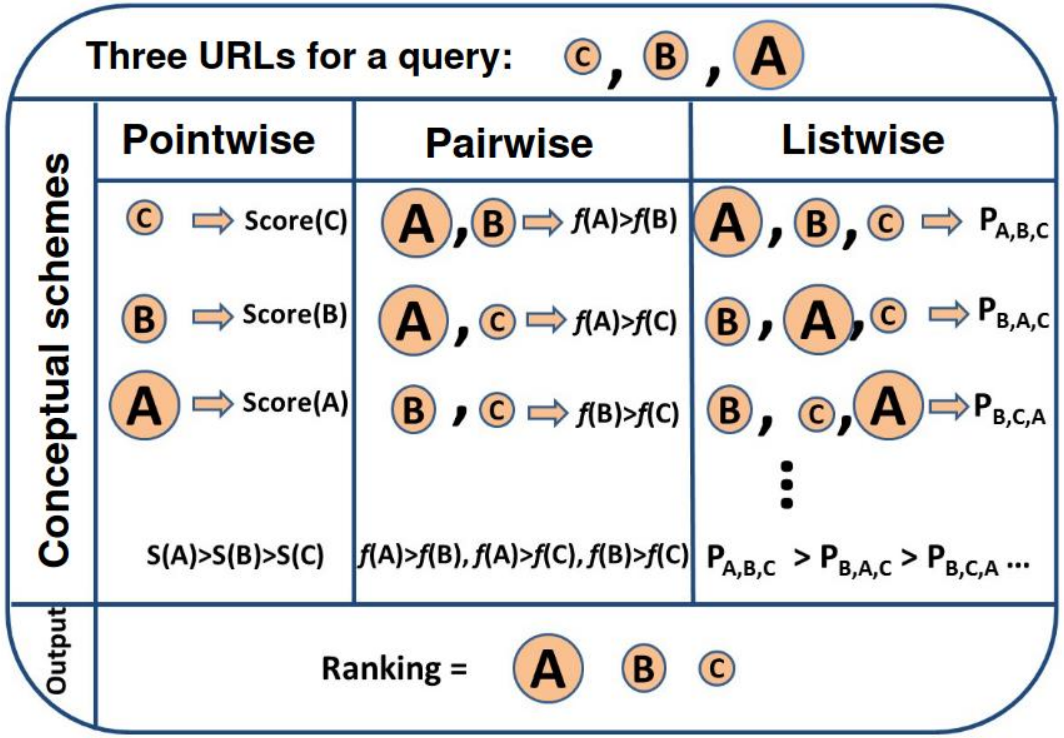
Retrieval Modeling and Pipeline

Modeling relevancy and do it fast

Three main modeling strategies

- **Pointwise**
- Pairwise
- Listwise

- And combinations of them



Three URLs for a query: C, B, A

	Pointwise	Pairwise	Listwise
Conceptual schemes	C ⇒ Score(C)	A, B ⇒ f(A)>f(B)	A, B, C ⇒ $P_{A,B,C}$
	B ⇒ Score(B)	A, C ⇒ f(A)>f(C)	B, A, C ⇒ $P_{B,A,C}$
	A ⇒ Score(A)	B, C ⇒ f(B)>f(C)	B, C, A ⇒ $P_{B,C,A}$
			⋮
	S(A)>S(B)>S(C)	f(A)>f(B), f(A)>f(C), f(B)>f(C)	$P_{A,B,C}$ > $P_{B,A,C}$ > $P_{B,C,A}$ ...
Output	Ranking = A B C		

https://medium.com/vptech/learning-to-rank-at-veepee-ed420fd828e5

Statistical Models

$$\text{score}(D, Q) = \sum_{\text{For each query term}} \boxed{\text{How important the term is}} \;\text{x}\; \boxed{\text{How often the term appear in the D}}$$

$$\text{score}(D, Q) = \sum_{\text{For each query term}} \boxed{\text{Inverted document frequency}} \;\text{x}\; \boxed{\text{Term frequency}}$$

TF-IDF

$$\text{score}(D, Q) = \sum_{i=1}^{n} log \frac{N}{n_t} \times \log(f(q_i, D) + 1)$$

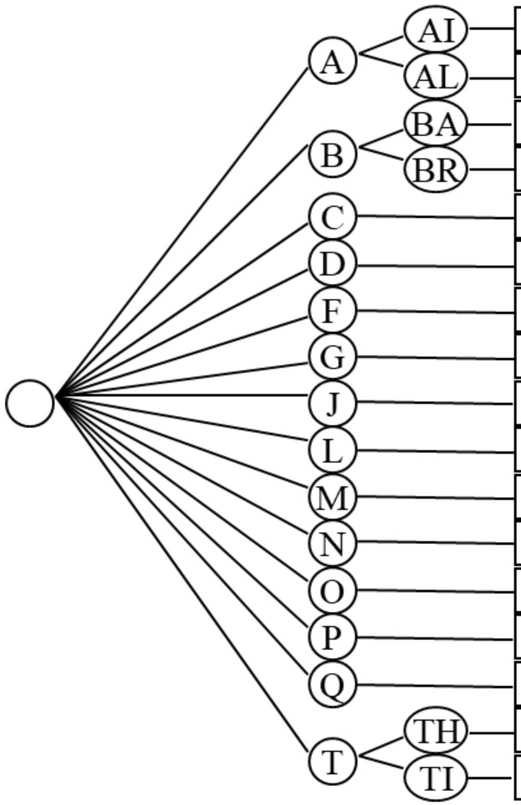
BM25

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

How to make it fast?

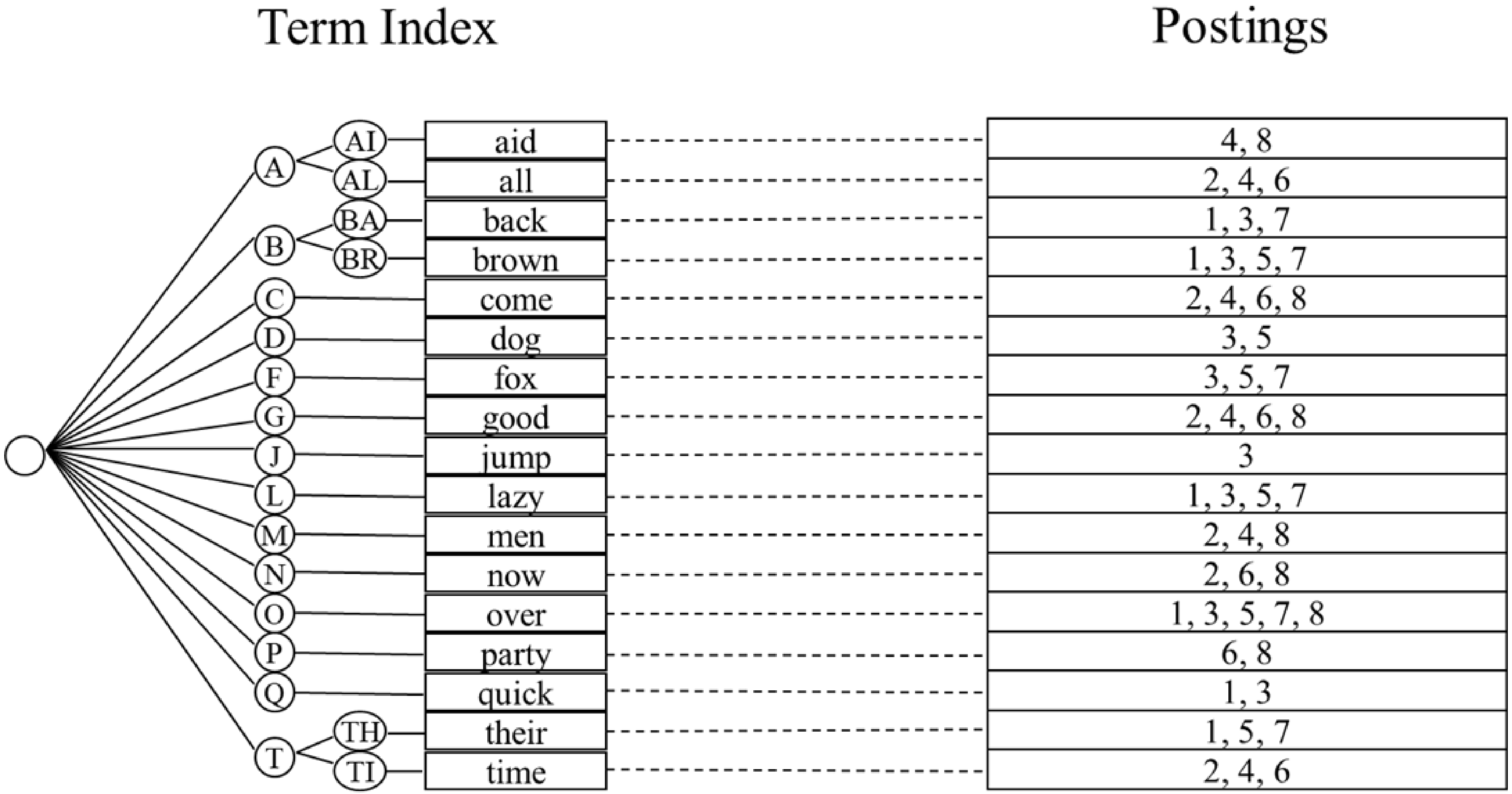
- “Fast” in responding to queries
- Better data structure
- Preprocess the data

Inverted Index



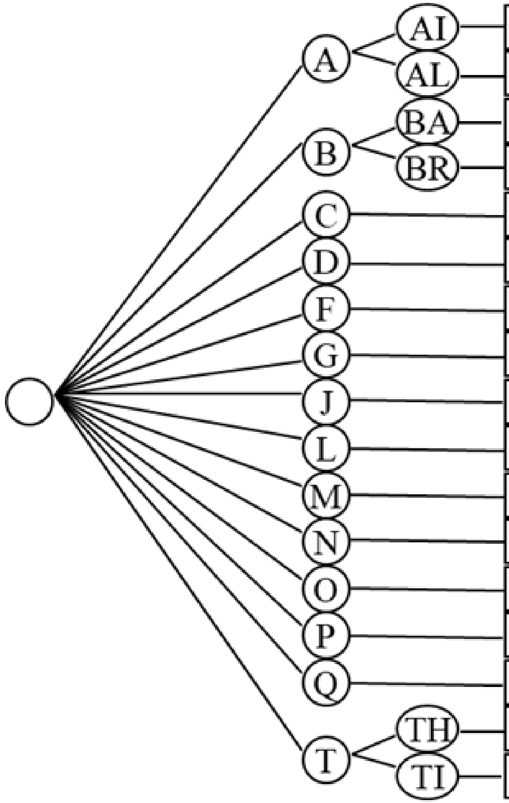
Term Index	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	Doc 6	Doc 7	Doc 8	Postings
aid				1				1	4, 8
all		1		1		1			2, 4, 6
back	1		1				1		1, 3, 7
brown	1		1		1		1		1, 3, 5, 7
come		1		1		1		1	2, 4, 6, 8
dog			1		1				3, 5
fox			1		1		1		3, 5, 7
good		1		1		1		1	2, 4, 6, 8
jump			1						3
lazy	1		1		1		1		1, 3, 5, 7
men		1		1				1	2, 4, 8
now		1				1		1	2, 6, 8
over	1		1		1		1	1	1, 3, 5, 7, 8
party						1		1	6, 8
quick	1		1						1, 3
their	1				1		1		1, 5, 7
time		1		1		1			2, 4, 6

Inverted Index



Term Index	Postings
aid	4, 8
all	2, 4, 6
back	1, 3, 7
brown	1, 3, 5, 7
come	2, 4, 6, 8
dog	3, 5
fox	3, 5, 7
good	2, 4, 6, 8
jump	3
lazy	1, 3, 5, 7
men	2, 4, 8
now	2, 6, 8
over	1, 3, 5, 7, 8
party	6, 8
quick	1, 3
their	1, 5, 7
time	2, 4, 6

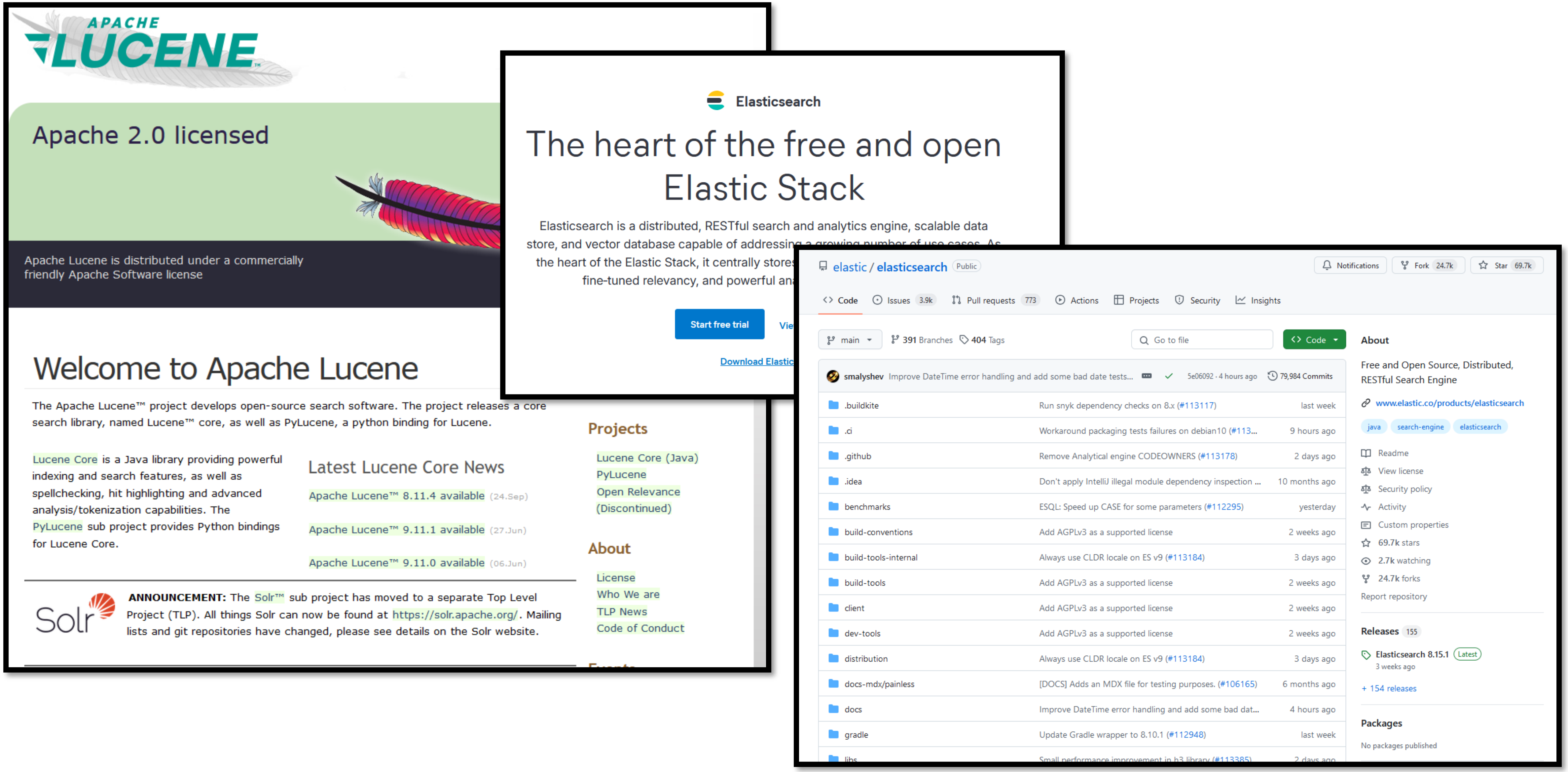
Inverted Index



Term Index	Postings
aid	4:0.86, 8:0.22
all	2:0.13, 4:0.59, 6:0.02
back	1:0.37, 3:0.34, 7:0.93
brown	1:0.79, 3:0.30, 5:0.19, 7:0.18
come	2:0.18, 4:0.56, 6:0.62, 8:0.02
dog	3:0.54, 5:0.13
fox	3:0.13, 5:0.09, 7:0.14
good	2:0.38, 4:0.85, 6:0.19, 8:0.01
jump	3:0.51
lazy	1:0.11, 3:0.17, 5:0.42, 7:0.84
men	2:0.15, 4:0.08, 8:0.66
now	2:0.82, 6:0.13, 8:0.38
over	1:0.22, 3:0.17, 5:0.37, 7:0.04, 8:0.11
party	6:0.49, 8:0.33
quick	1:0.19, 3:0.54
their	1:0.58, 5:0.17, 7:0.23
time	2:0.83, 4:0.22, 6:0.28

Two-Stage System

- Offline preprocessing and indexing
    - Define retrieval unit
    - Tokenization
    - Build the inverted index
- Online query serving
    - Traverse the inverted index and score it



Apache 2.0 licensed

Apache Lucene is distributed under a commercially friendly Apache Software license

Welcome to Apache Lucene

The Apache Lucene™ project develops open-source search software. The project releases a core search library, named Lucene™ core, as well as PyLucene, a python binding for Lucene.

Lucene Core is a Java library providing powerful indexing and search features, as well as spellchecking, hit highlighting and advanced analysis/tokenization capabilities. The PyLucene sub project provides Python bindings for Lucene Core.

Latest Lucene Core News

Apache Lucene™ 8.11.4 available (24.Sep)

Apache Lucene™ 9.11.1 available (27.Jun)

Apache Lucene™ 9.11.0 available (06.Jun)

ANNOUNCEMENT: The Solr™ sub project has moved to a separate Top Level Project (TLP). All things Solr can now be found at https://solr.apache.org/. Mailing lists and git repositories have changed, please see details on the Solr website.

Projects

Lucene Core (Java)
PyLucene
Open Relevance (Discontinued)

About

License
Who We are
TLP News
Code of Conduct

Events

The heart of the free and open Elastic Stack

Elasticsearch is a distributed, RESTful search and analytics engine, scalable data store, and vector database capable of addressing a growing number of use cases. As the heart of the Elastic Stack, it centrally stores
fine-tuned relevancy, and powerful an

smalyshev Improve DateTime error handling and add some bad date tests... ✓ 5e06092 · 4 hours ago 79,984 Commits

.buildkite	Run snyk dependency checks on 8.x (#113117)	last week
.ci	Workaround packaging tests failures on debian10 (#113...	9 hours ago
.github	Remove Analytical engine CODEOWNERS (#113178)	2 days ago
.idea	Don't apply IntelliJ illegal module dependency inspection ...	10 months ago
benchmarks	ESQL: Speed up CASE for some parameters (#112295)	yesterday
build-conventions	Add AGPLv3 as a supported license	2 weeks ago
build-tools-internal	Always use CLDR locale on ES v9 (#113184)	3 days ago
build-tools	Add AGPLv3 as a supported license	2 weeks ago
client	Add AGPLv3 as a supported license	2 weeks ago
dev-tools	Add AGPLv3 as a supported license	2 weeks ago
distribution	Always use CLDR locale on ES v9 (#113184)	3 days ago
docs-mdx/painless	[DOCS] Adds an MDX file for testing purposes. (#106165)	6 months ago
docs	Improve DateTime error handling and add some bad dat...	4 hours ago
gradle	Update Gradle wrapper to 8.10.1 (#112948)	last week
libs	Small performance improvement in h3 library (#113385)	2 days ago

About

Free and Open Source, Distributed, RESTful Search Engine

www.elastic.co/products/elasticsearch

69.7k stars
2.7k watching
24.7k forks

Releases 155

Elasticsearch 8.15.1 Latest
3 weeks ago

+ 154 releases

Packages

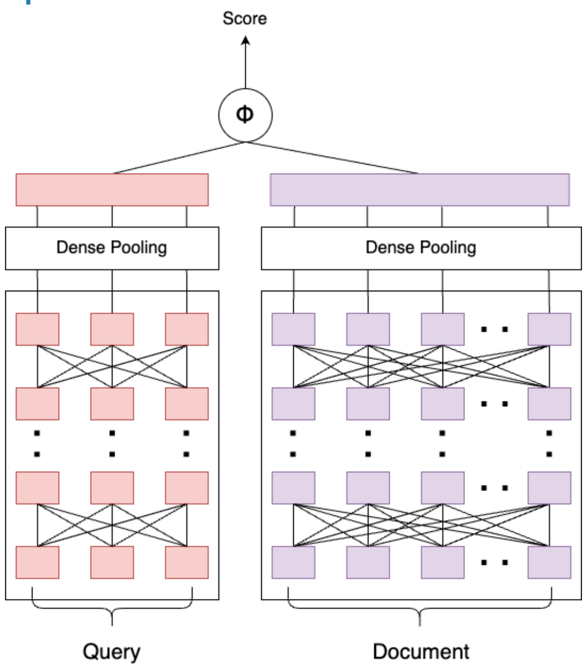
No packages published

Can we go beyond surface forms?

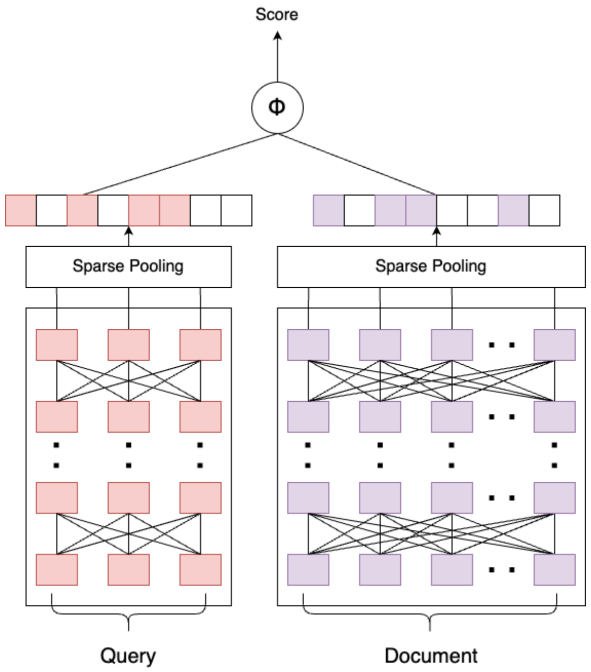
neural language models

Bi-Encoder

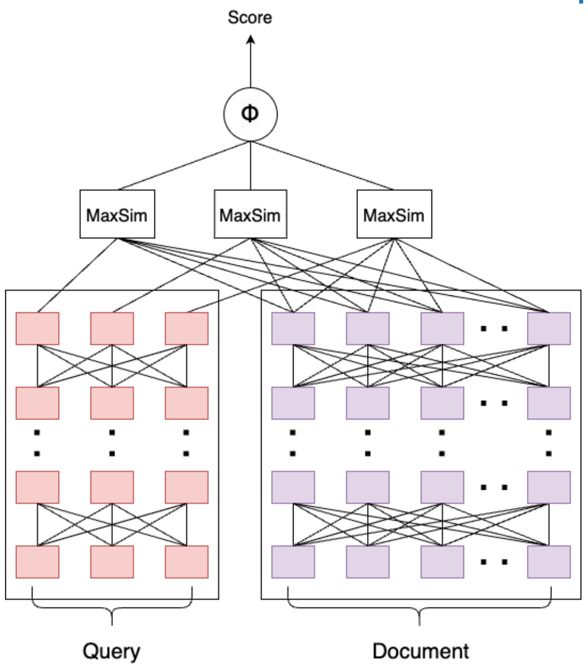
Cross Encoder



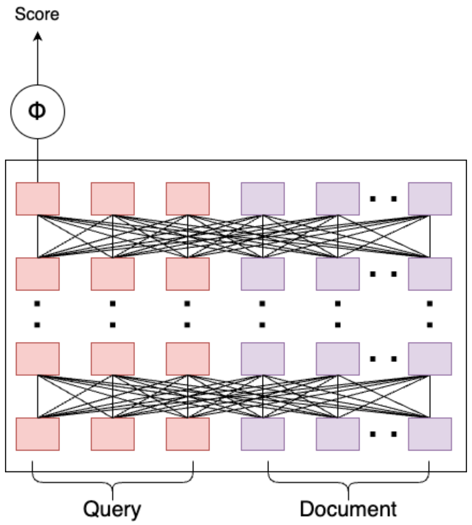
One Dense Vector
Per Sequence
e.g., DPR



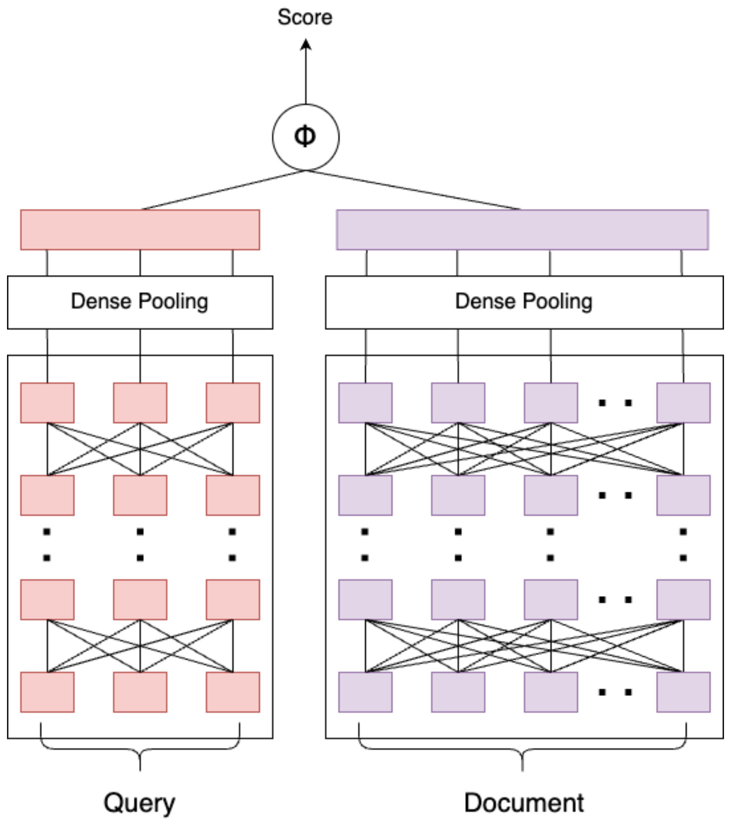
One **Sparse** Vector
Per Sequence
e.g., SPLADE



Multiple Dense Vectors
Per Sequence
e.g., ColBERT

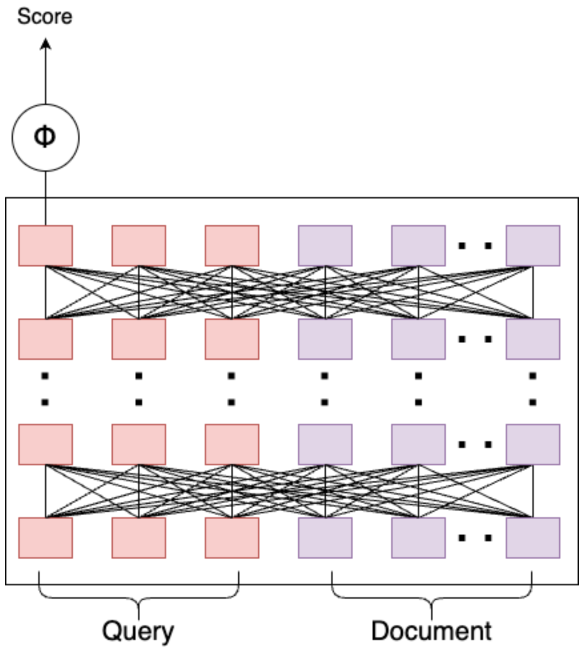


Joint Encoder
e.g., monoBERT



Online

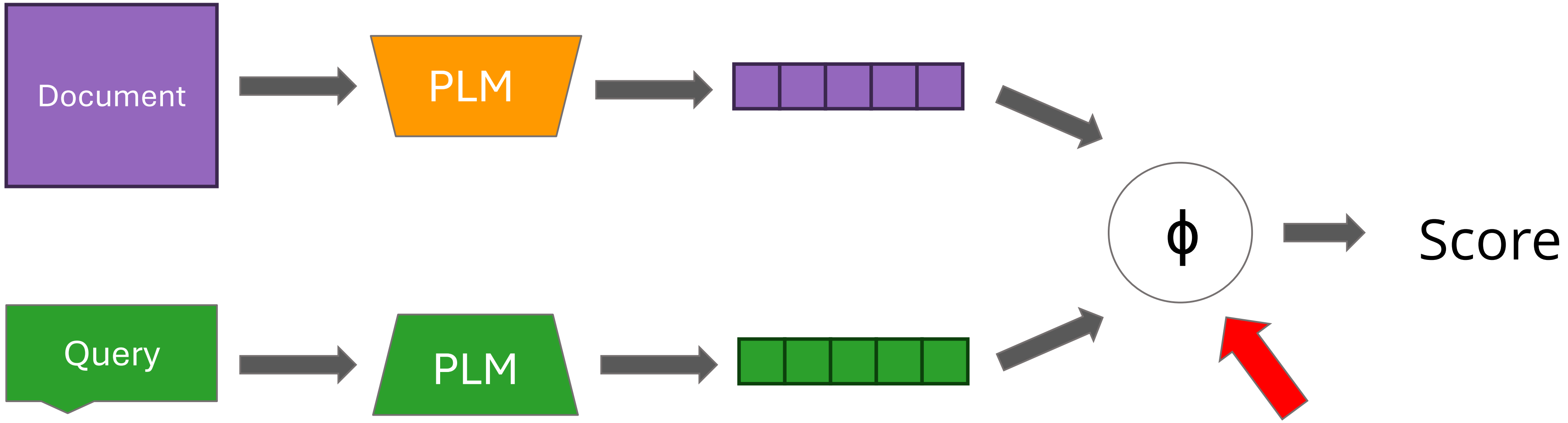
Offline



Both Online

Separate query and document processing

One Vector per Query, One Vector per Document



Document

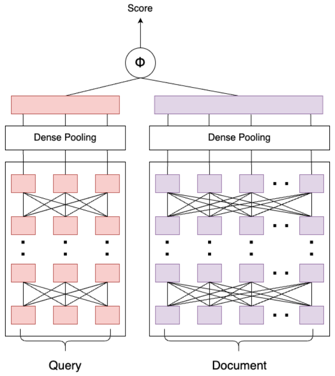
PLM

Query

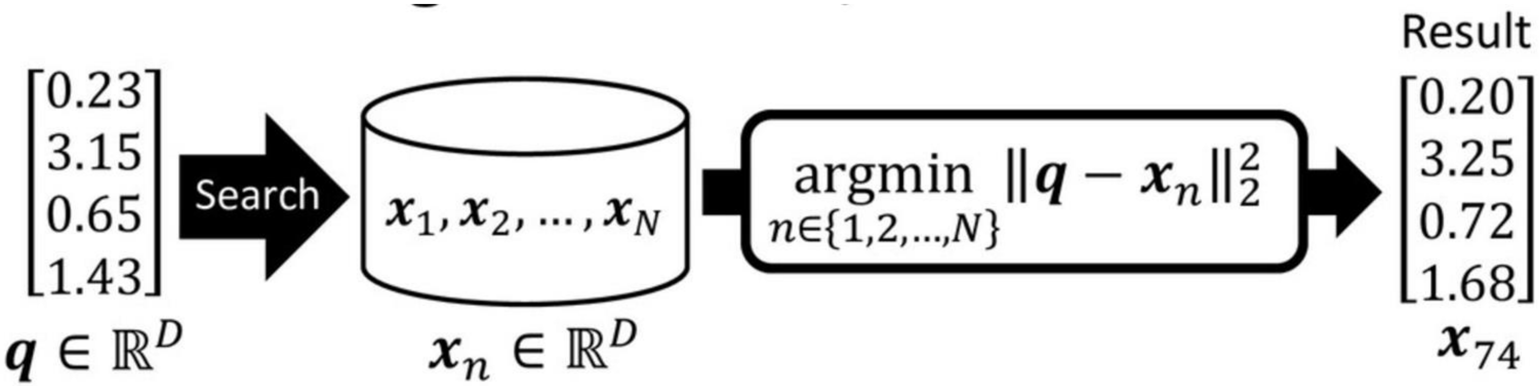
PLM

ϕ

Score



Nearest Vectors aka Neighbors

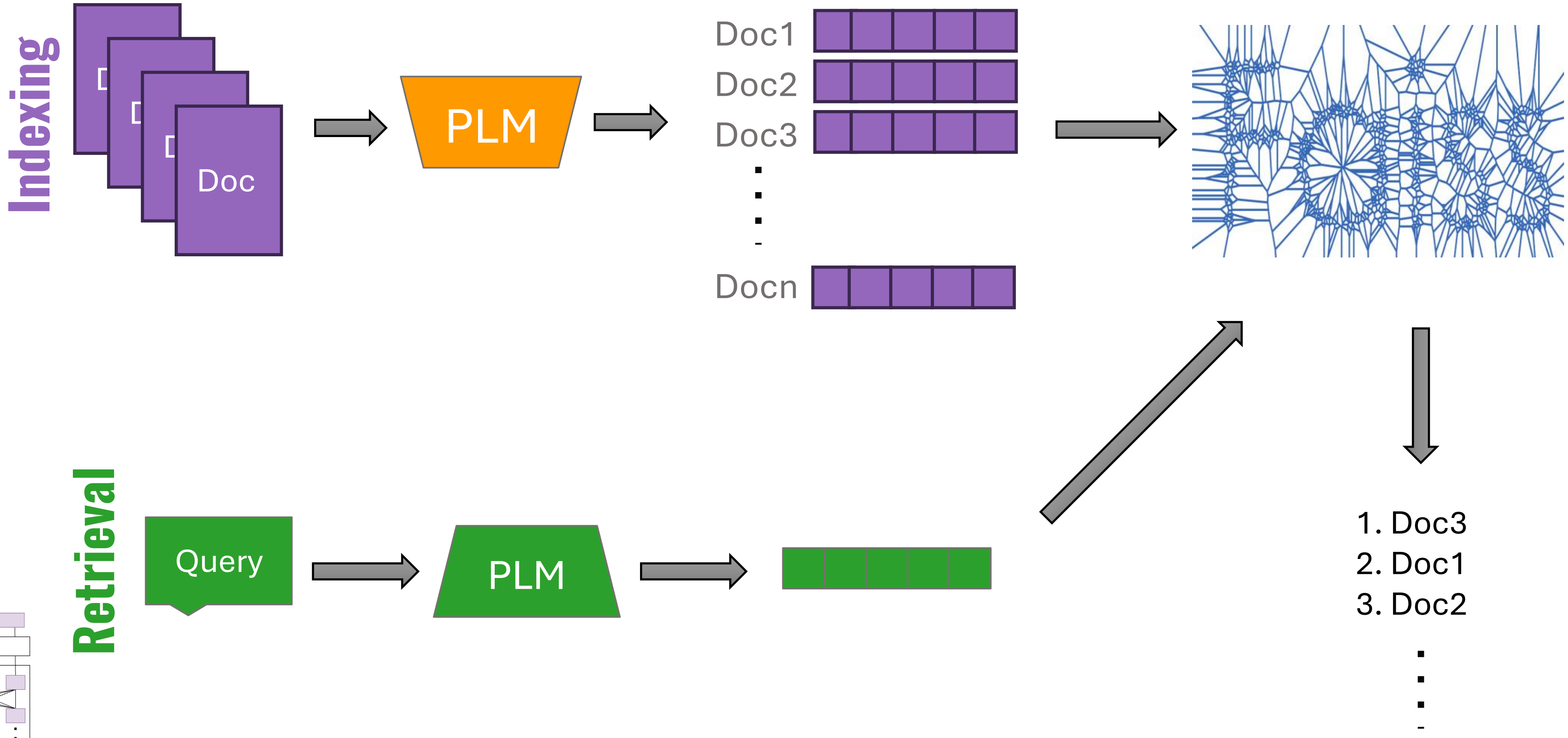

$\begin{bmatrix}0.23\\3.15\\0.65\\1.43\end{bmatrix}$
$\boldsymbol{q} \in \mathbb{R}^D$
Search
$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N$
$\boldsymbol{x}_n \in \mathbb{R}^D$
$\underset{n \in \{1,2,\dots,N\}}{\operatorname{argmin}} \|\boldsymbol{q} - \boldsymbol{x}_n\|_2^2$
Result
$\begin{bmatrix}0.20\\3.25\\0.72\\1.68\end{bmatrix}$
$\boldsymbol{x}_{74}$

- Linear Search
  - Slow (scales linearly in size of document collection)

- Approximate Methods (e.g., Product Quantization) → **ANN**
  - Faster Search

- Runtime Efficiency vs Effectiveness

DPR Indexing and Retrieval



Indexing

Doc

PLM

Doc1

Doc2

Doc3

⋮

Docn

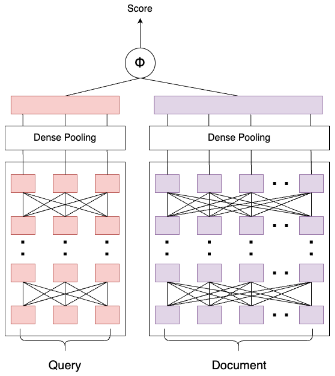
Retrieval

Query

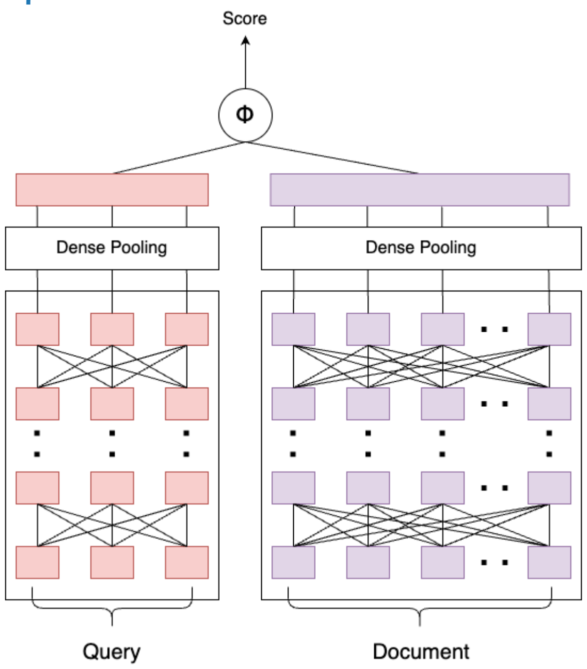
PLM

1. Doc3
2. Doc1
3. Doc2

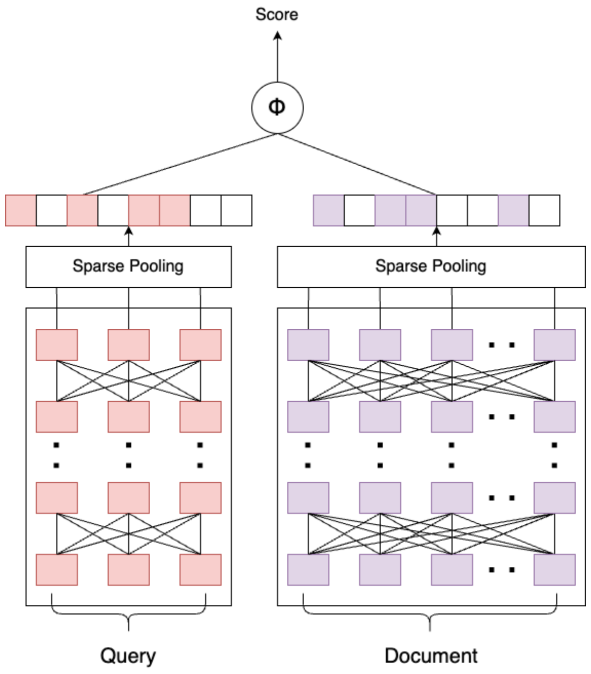
⋮



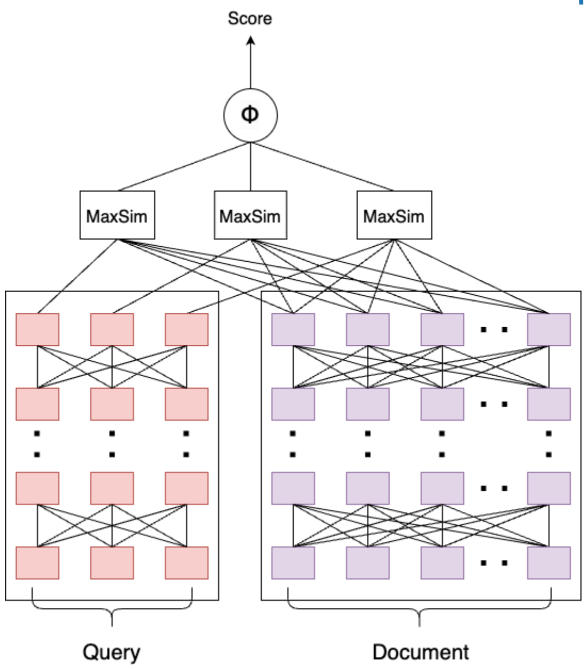
Bi-Encoder



One Dense Vector
Per Sequence
e.g., DPR

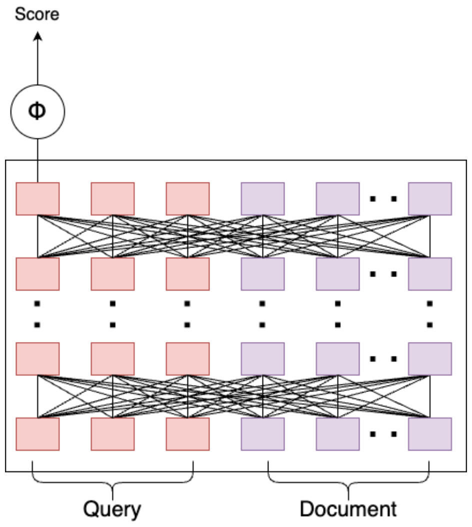


One **Sparse** Vector
Per Sequence
e.g., SPLADE



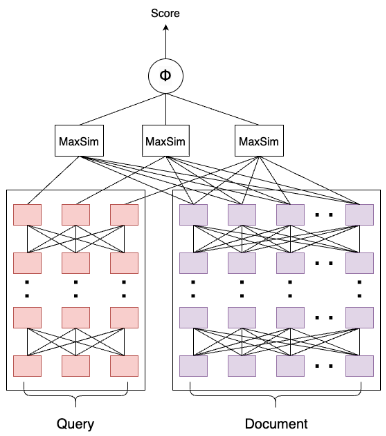
Multiple Dense Vectors
Per Sequence
e.g., ColBERT

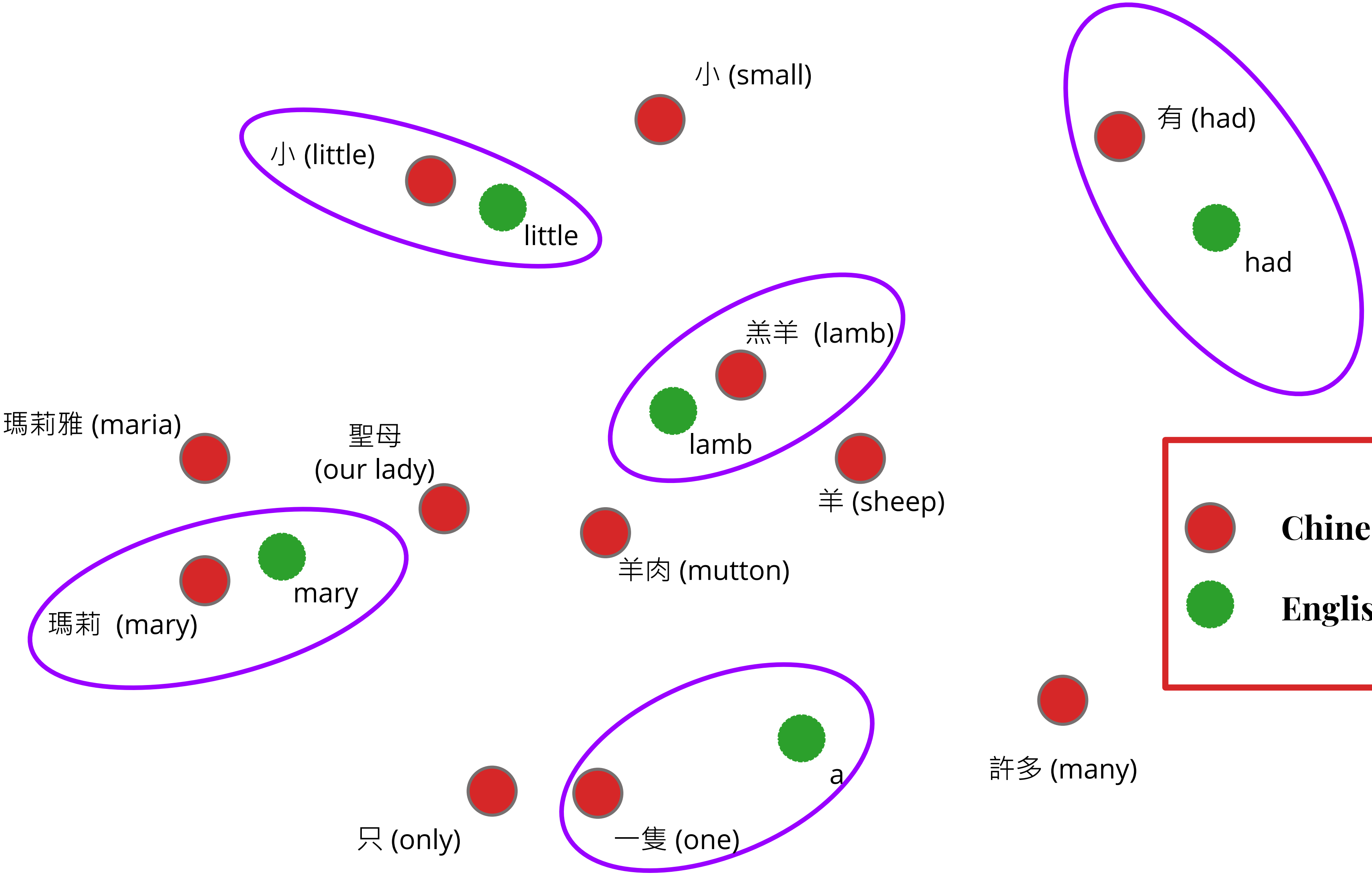
Cross Encoder



Joint Encoder
e.g., monoBERT

One Vector per Term: MaxSim

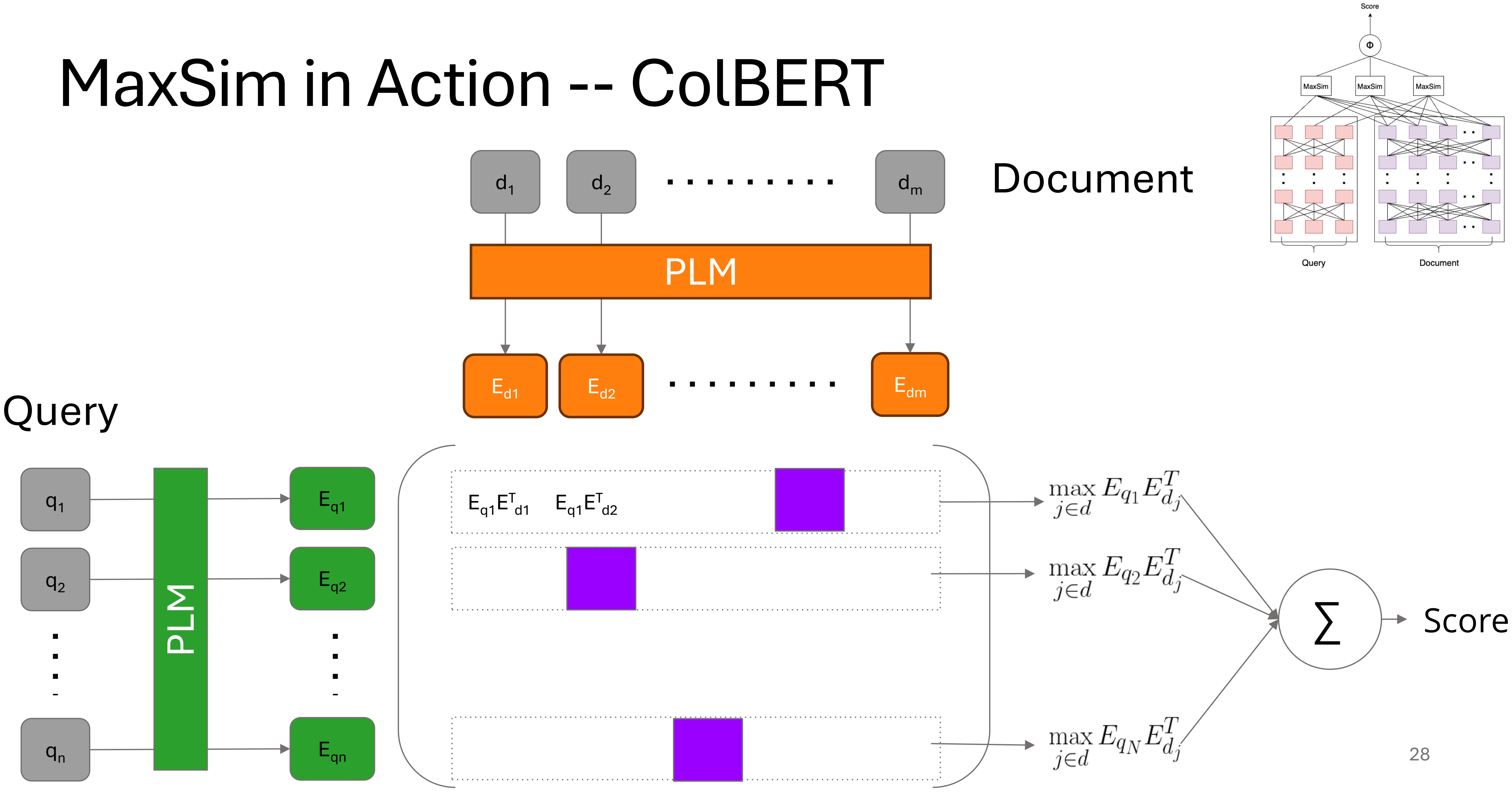




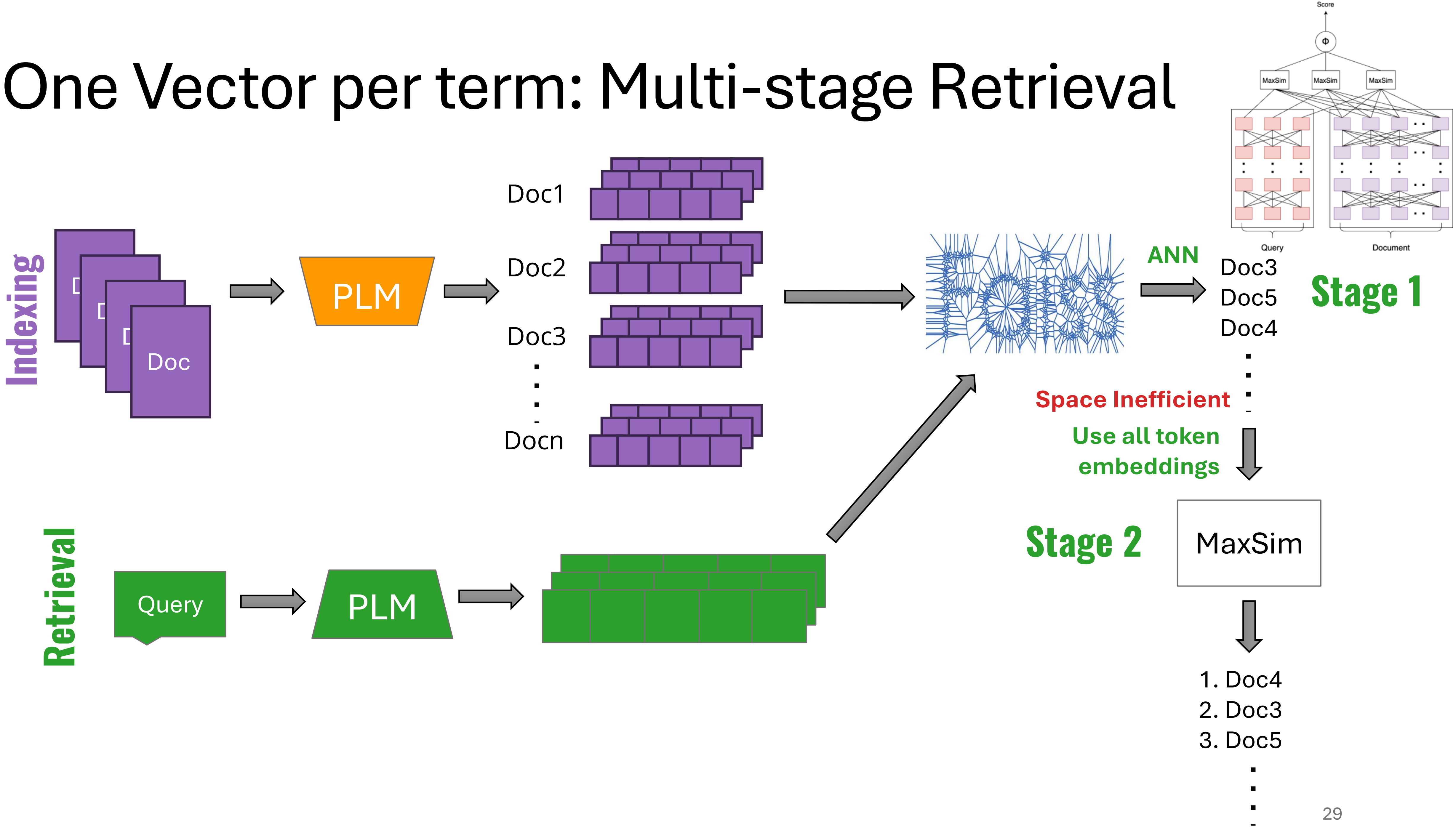
🔴 **Chinese Document term embedding**

🟢 **English Query term embedding**

MaxSim in Action -- ColBERT



One Vector per term: Multi-stage Retrieval



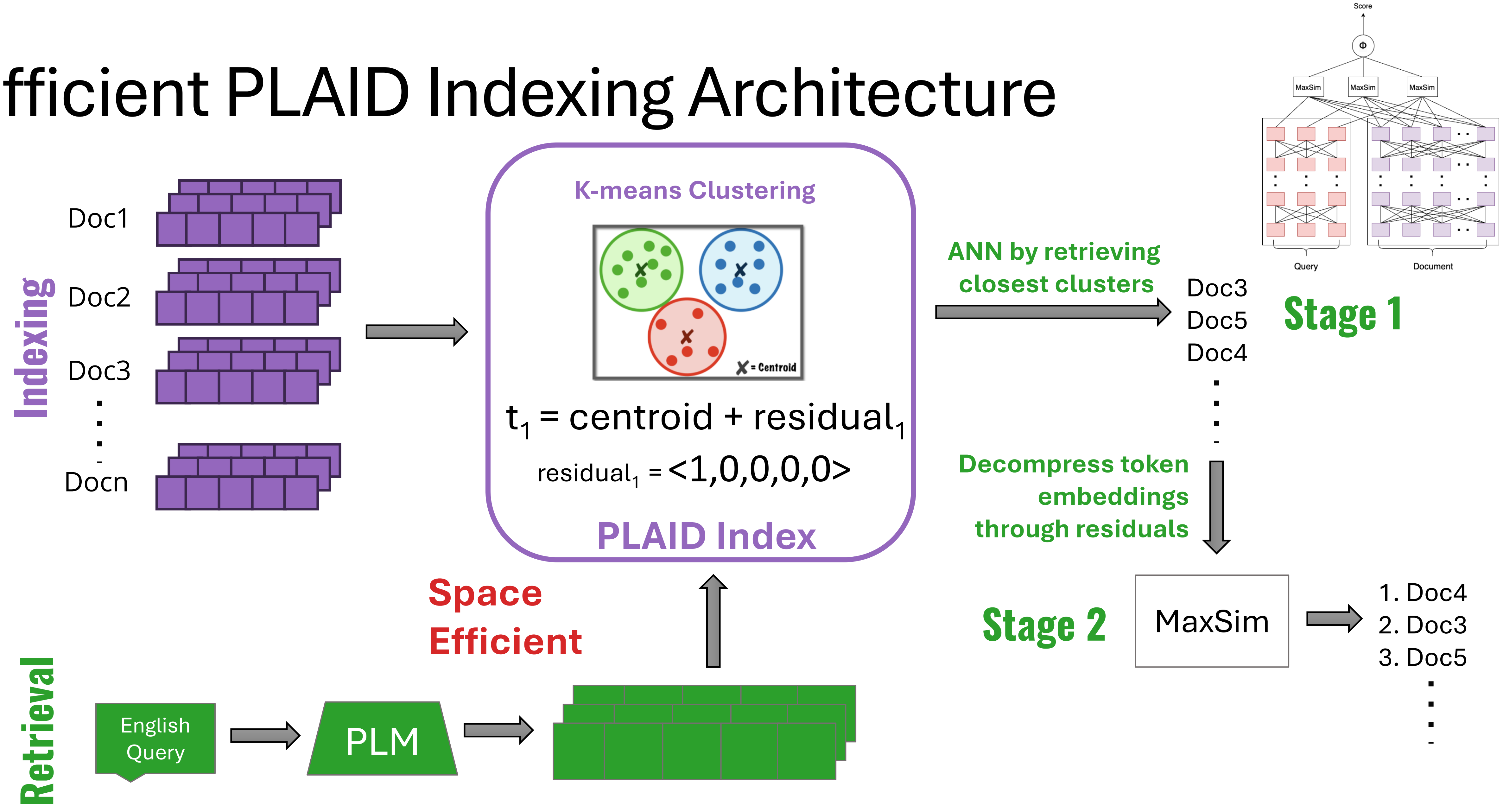
Stage 1

Space Inefficient

Use all token embeddings

Stage 2

Efficient PLAID Indexing Architecture



Indexing

Doc1

Doc2

Doc3

⋮

Docn

K-means Clustering

✗ = Centroid

t_1 = centroid + residual_1

residual_1 = <1,0,0,0,0>

PLAID Index

ANN by retrieving closest clusters

Doc3
Doc5
Doc4
⋮

Stage 1

Decompress token embeddings through residuals

Stage 2

MaxSim

1. Doc4
2. Doc3
3. Doc5

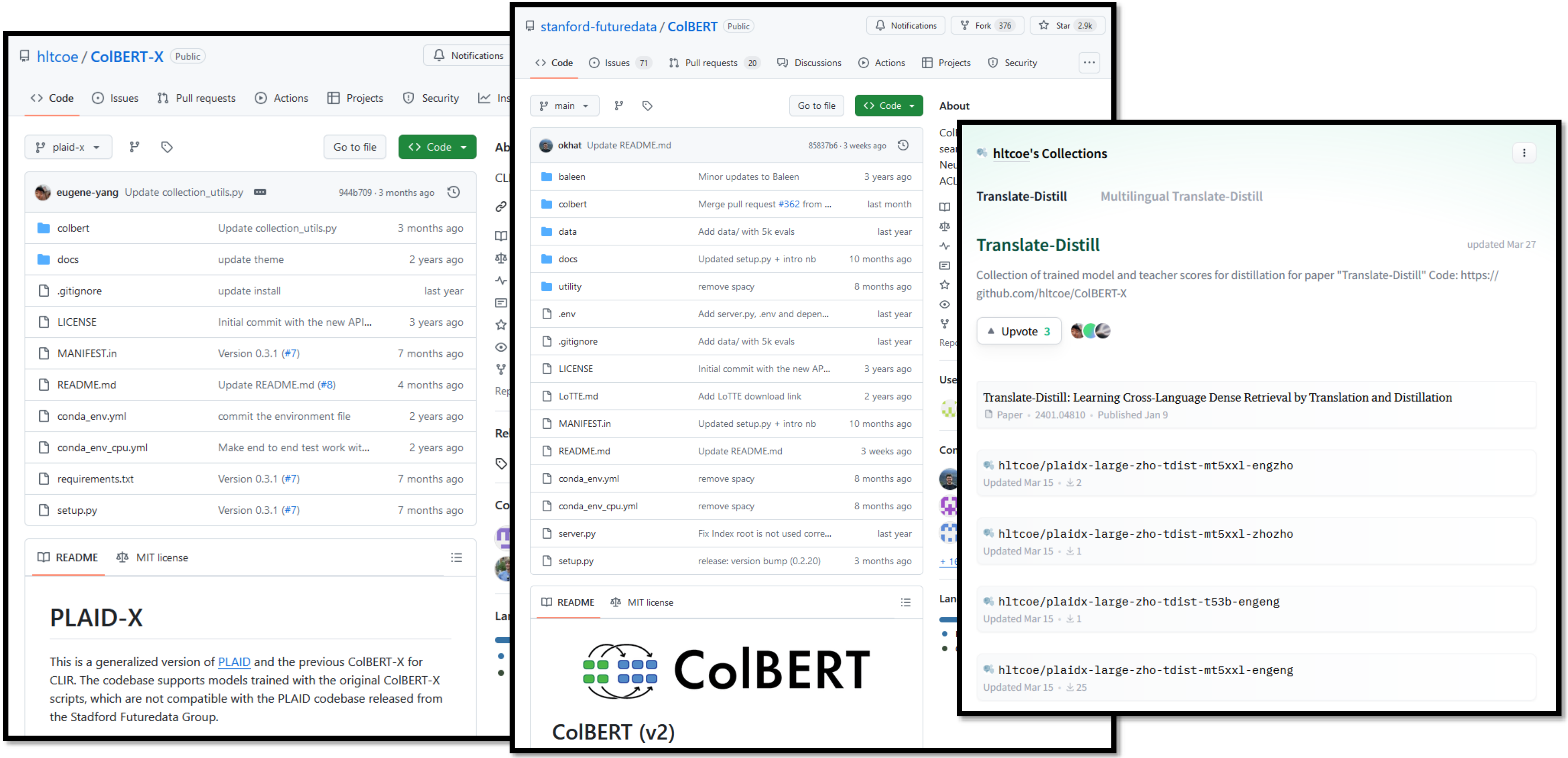
⋮

Space Efficient

Retrieval

English Query

PLM



hltcoe / ColBERT-X Public

Notifications

Code · Issues · Pull requests · Actions · Projects · Security · Ins

plaid-x · Go to file · Code

eugene-yang Update collection_utils.py · 944b709 · 3 months ago

colbert	Update collection_utils.py	3 months ago
docs	update theme	2 years ago
.gitignore	update install	last year
LICENSE	Initial commit with the new API...	3 years ago
MANIFEST.in	Version 0.3.1 (#7)	7 months ago
README.md	Update README.md (#8)	4 months ago
conda_env.yml	commit the environment file	2 years ago
conda_env_cpu.yml	Make end to end test work wit...	2 years ago
requirements.txt	Version 0.3.1 (#7)	7 months ago
setup.py	Version 0.3.1 (#7)	7 months ago

README · MIT license

PLAID-X

This is a generalized version of PLAID and the previous ColBERT-X for CLIR. The codebase supports models trained with the original ColBERT-X scripts, which are not compatible with the PLAID codebase released from the Stadford Futuredata Group.

stanford-futuredata / ColBERT Public

Notifications · Fork 376 · Star 2.9k

Code · Issues 71 · Pull requests 20 · Discussions · Actions · Projects · Security

main · Go to file · Code

okhat Update README.md · 85837b6 · 3 weeks ago

baleen	Minor updates to Baleen	3 years ago
colbert	Merge pull request #362 from ...	last month
data	Add data/ with 5k evals	last year
docs	Updated setup.py + intro nb	10 months ago
utility	remove spacy	8 months ago
.env	Add server.py, .env and depen...	last year
.gitignore	Add data/ with 5k evals	last year
LICENSE	Initial commit with the new AP...	3 years ago
LoTTE.md	Add LoTTE download link	2 years ago
MANIFEST.in	Updated setup.py + intro nb	10 months ago
README.md	Update README.md	3 weeks ago
conda_env.yml	remove spacy	8 months ago
conda_env_cpu.yml	remove spacy	8 months ago
server.py	Fix Index root is not used corre...	last year
setup.py	release: version bump (0.2.20)	3 months ago

README · MIT license

ColBERT

ColBERT (v2)

About

hltcoe's Collections

Translate-Distill · Multilingual Translate-Distill

Translate-Distill

updated Mar 27

Collection of trained model and teacher scores for distillation for paper "Translate-Distill" Code: https://github.com/hltcoe/ColBERT-X

Upvote 3

Translate-Distill: Learning Cross-Language Dense Retrieval by Translation and Distillation
Paper · 2401.04810 · Published Jan 9

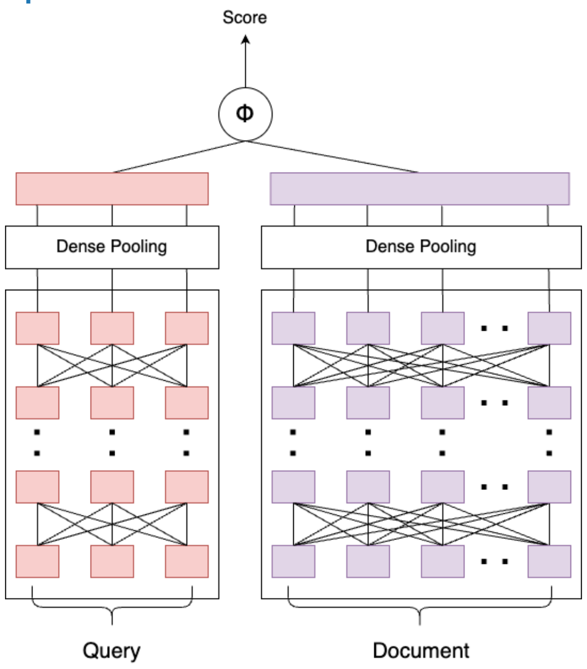
hltcoe/plaidx-large-zho-tdist-mt5xxl-engzho
Updated Mar 15 · 2

hltcoe/plaidx-large-zho-tdist-mt5xxl-zhozho
Updated Mar 15 · 1

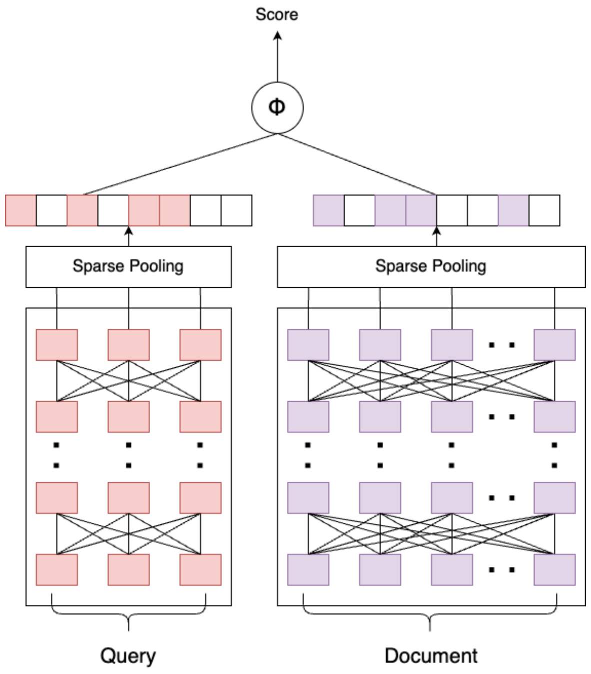
hltcoe/plaidx-large-zho-tdist-t53b-engeng
Updated Mar 15 · 1

hltcoe/plaidx-large-zho-tdist-mt5xxl-engeng
Updated Mar 15 · 25

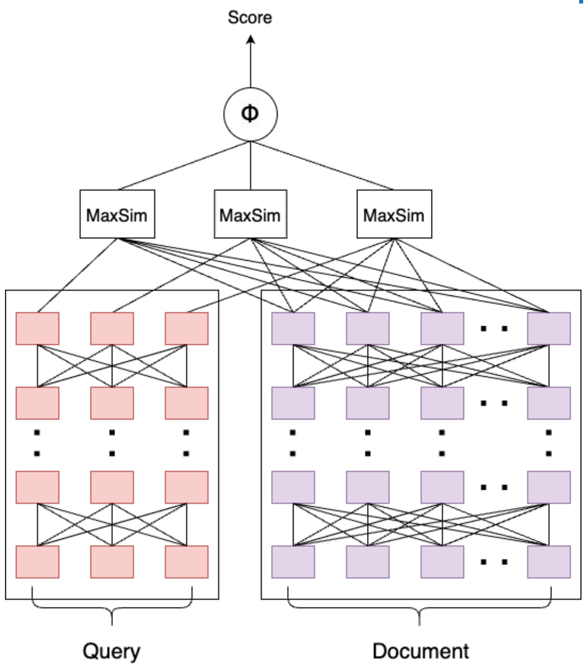
Bi-Encoder



One Dense Vector
Per Sequence
e.g., DPR

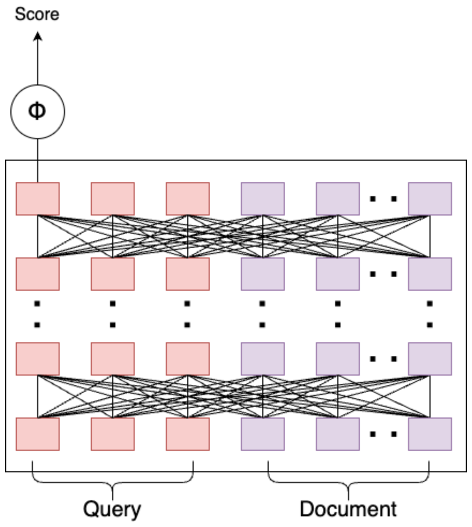


One **Sparse** Vector
Per Sequence
e.g., SPLADE



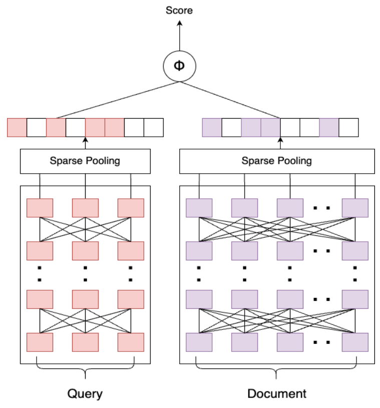
Multiple Dense Vectors
Per Sequence
e.g., ColBERT

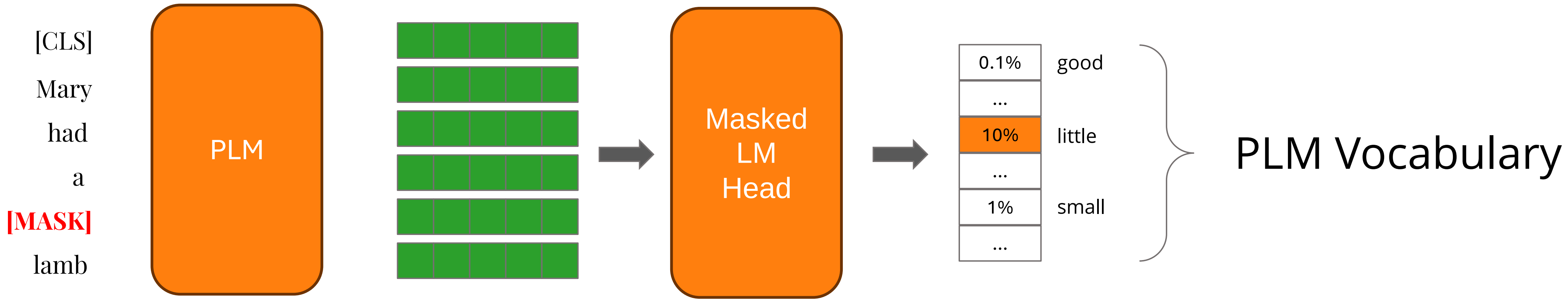
Cross Encoder



Joint Encoder
e.g., monoBERT

High-dimensional Vector: Masked LM





[CLS]
Mary
had
a
[MASK]
lamb

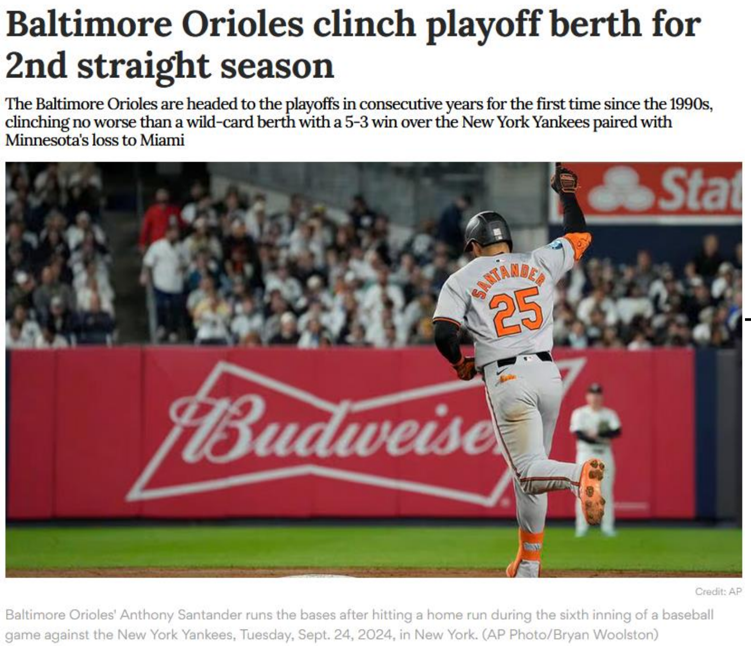
PLM

Masked LM Head

0.1%	good
…	
10%	little
…	
1%	small
…	

PLM Vocabulary

SPLADE

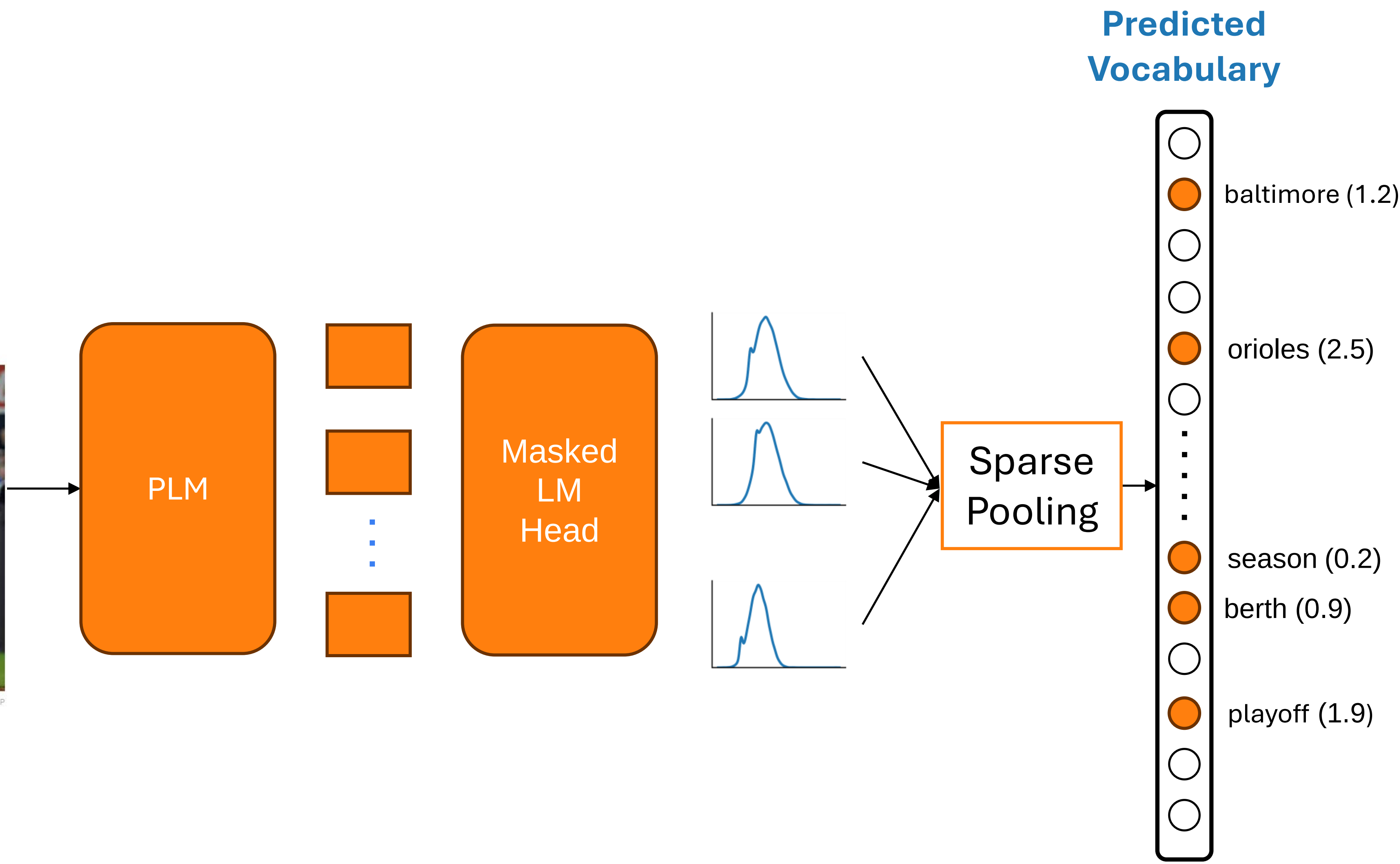


Baltimore Orioles clinch playoff berth for 2nd straight season

The Baltimore Orioles are headed to the playoffs in consecutive years for the first time since the 1990s, clinching no worse than a wild-card berth with a 5-3 win over the New York Yankees paired with Minnesota's loss to Miami

Credit: AP

Baltimore Orioles' Anthony Santander runs the bases after hitting a home run during the sixth inning of a baseball game against the New York Yankees, Tuesday, Sept. 24, 2024, in New York. (AP Photo/Bryan Woolston)



Predicted Vocabulary

PLM

Masked LM Head

Sparse Pooling

baltimore (1.2)

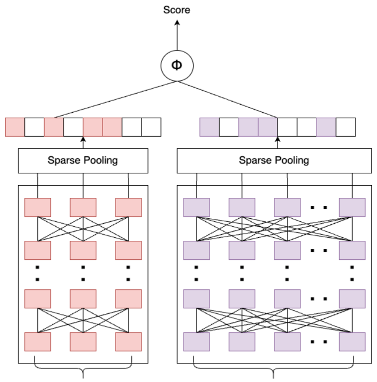
orioles (2.5)

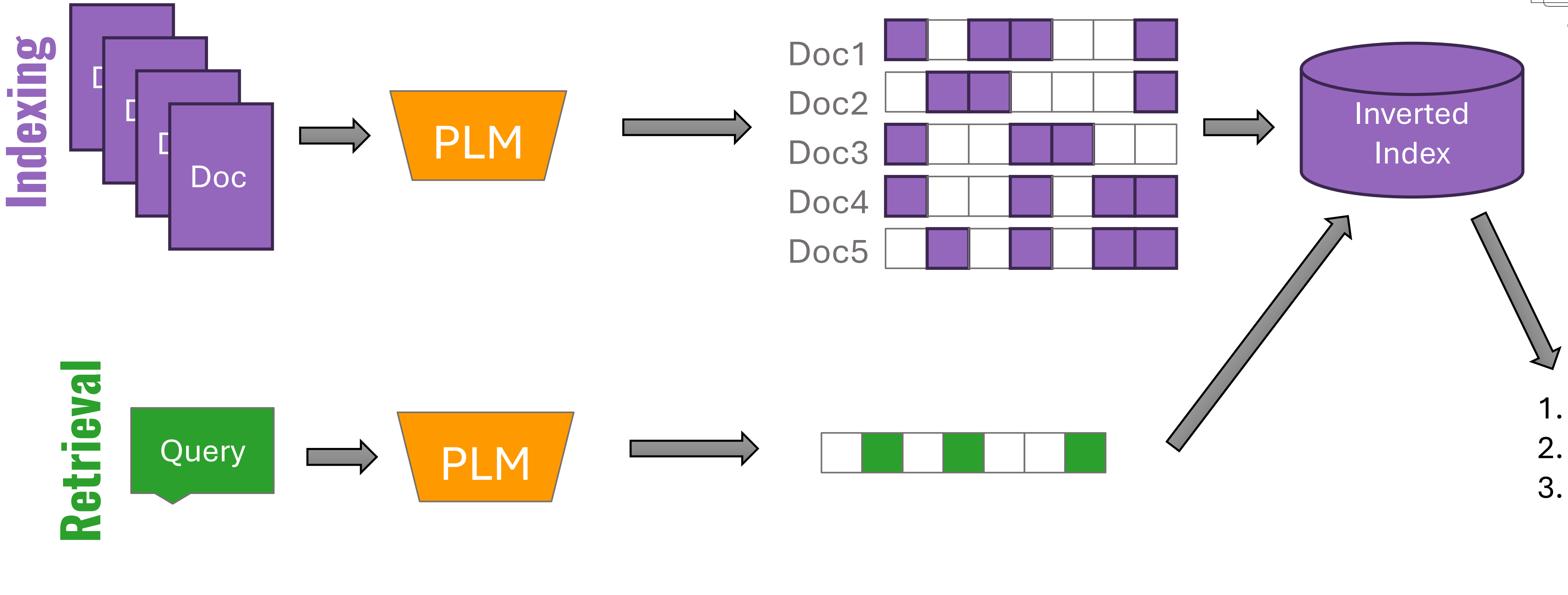
season (0.2)

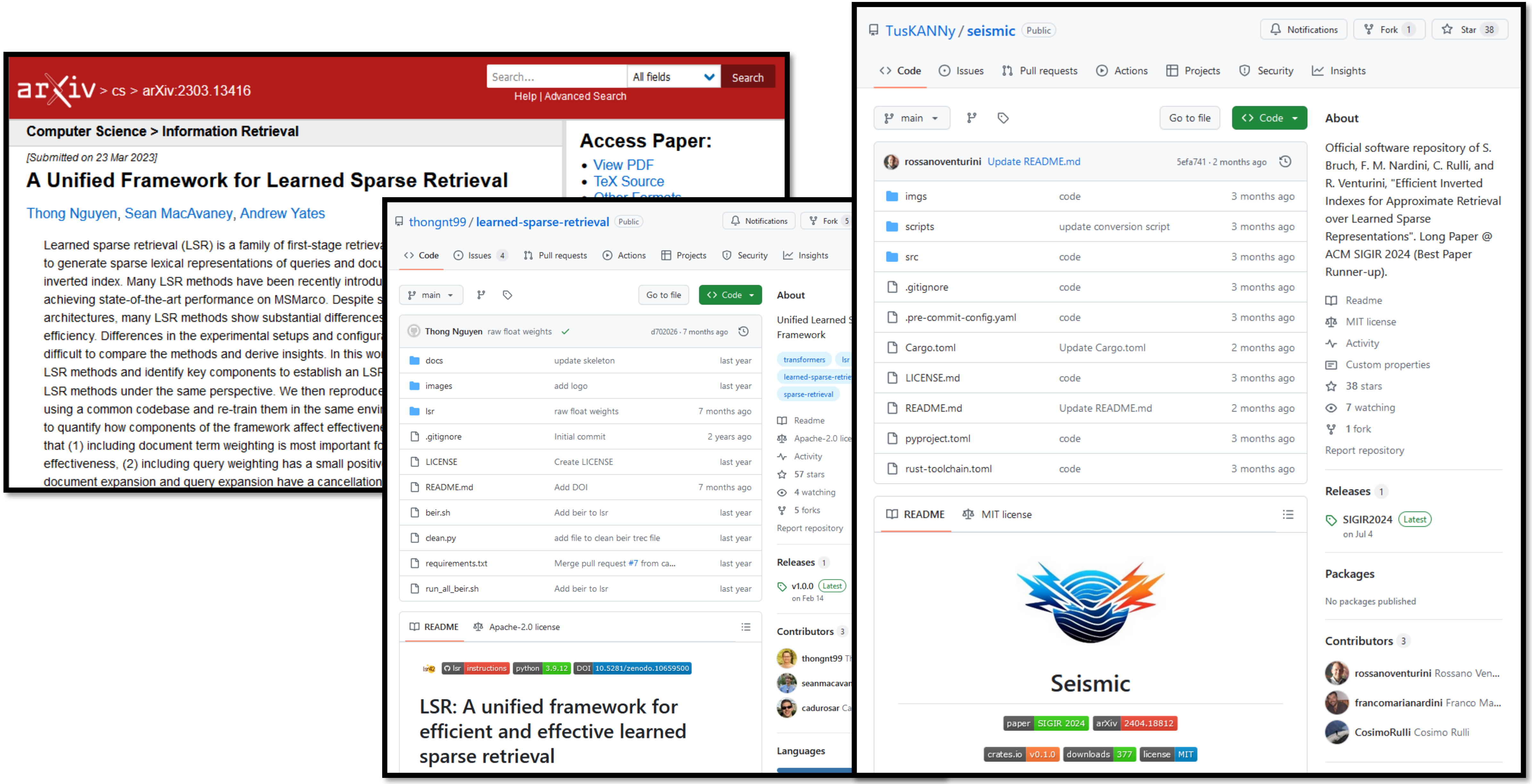
berth (0.9)

playoff (1.9)

SPLADE Search Pipeline







Computer Science > Information Retrieval

[Submitted on 23 Mar 2023]

A Unified Framework for Learned Sparse Retrieval

Thong Nguyen, Sean MacAvaney, Andrew Yates

Learned sparse retrieval (LSR) is a family of first-stage retrieva
to generate sparse lexical representations of queries and docu
inverted index. Many LSR methods have been recently introdu
achieving state-of-the-art performance on MSMarco. Despite s
architectures, many LSR methods show substantial differences
efficiency. Differences in the experimental setups and configura
difficult to compare the methods and derive insights. In this wo
LSR methods and identify key components to establish an LSR
LSR methods under the same perspective. We then reproduce
using a common codebase and re-train them in the same envir
to quantify how components of the framework affect effectivene
that (1) including document term weighting is most important fo
effectiveness, (2) including query weighting has a small positiv
document expansion and query expansion have a cancellation

Access Paper:

- View PDF
- TeX Source
- Other Formats

Thong Nguyen raw float weights ✓ d702026 · 7 months ago

docs	update skeleton	last year
images	add logo	last year
lsr	raw float weights	7 months ago
.gitignore	Initial commit	2 years ago
LICENSE	Create LICENSE	last year
README.md	Add DOI	7 months ago
beir.sh	Add beir to lsr	last year
clean.py	add file to clean beir trec file	last year
requirements.txt	Merge pull request #7 from ca...	last year
run_all_beir.sh	Add beir to lsr	last year

LSR: A unified framework for efficient and effective learned sparse retrieval

About

Unified Learned S
Framework

rossanoventurini Update README.md 5efa741 · 2 months ago

imgs	code	3 months ago
scripts	update conversion script	3 months ago
src	code	3 months ago
.gitignore	code	3 months ago
.pre-commit-config.yaml	code	3 months ago
Cargo.toml	Update Cargo.toml	2 months ago
LICENSE.md	code	3 months ago
README.md	Update README.md	2 months ago
pyproject.toml	code	3 months ago
rust-toolchain.toml	code	3 months ago

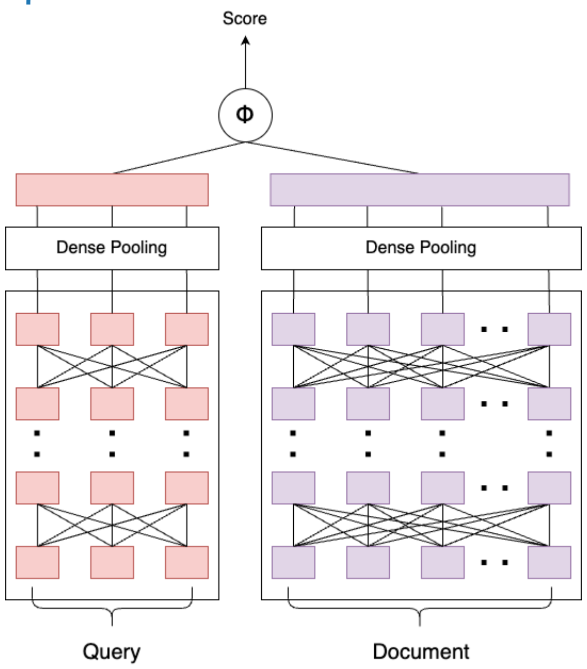
Seismic

About

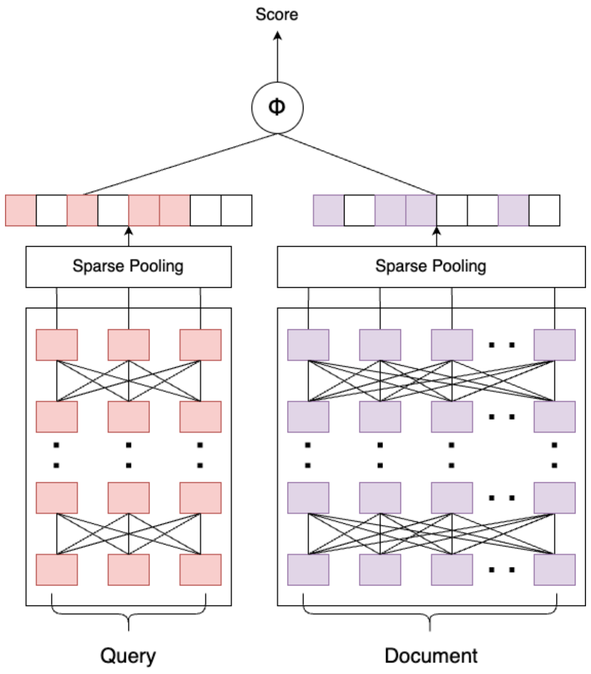
Official software repository of S. Bruch, F. M. Nardini, C. Rulli, and R. Venturini, "Efficient Inverted Indexes for Approximate Retrieval over Learned Sparse Representations". Long Paper @ ACM SIGIR 2024 (Best Paper Runner-up).

Bi-Encoder

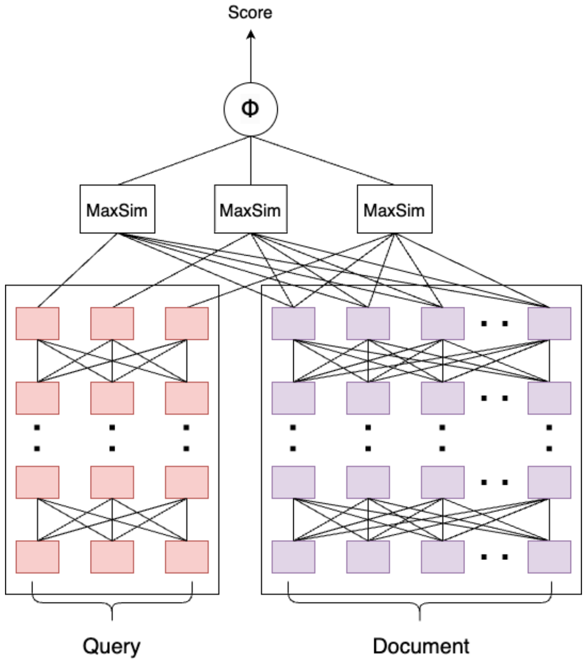
Cross Encoder



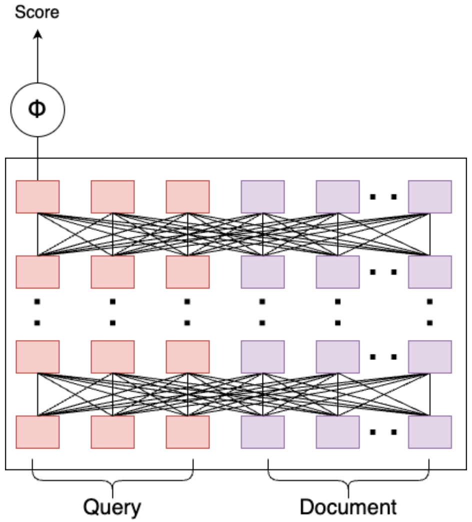
One Dense Vector
Per Sequence
e.g., DPR



One **Sparse** Vector
Per Sequence
e.g., SPLADE

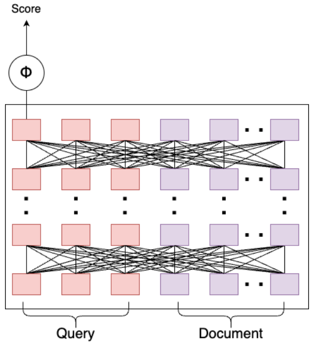


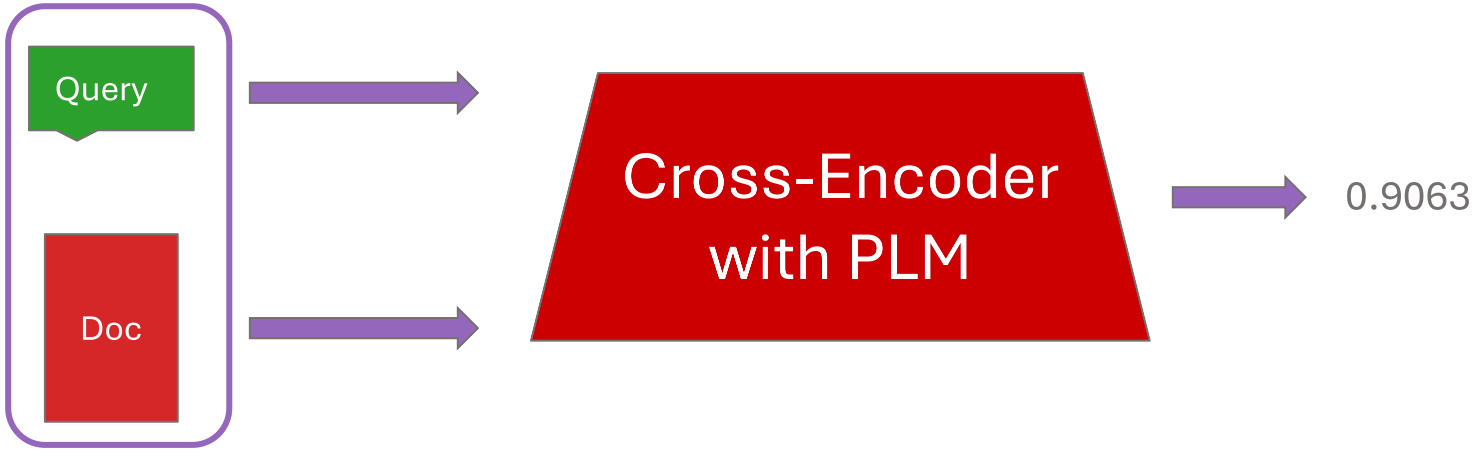
Multiple Dense Vectors
Per Sequence
e.g., ColBERT



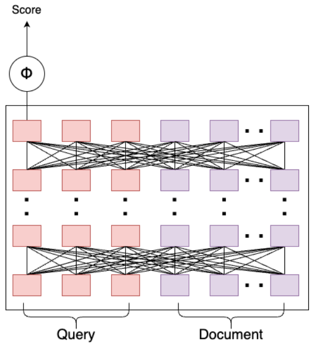
Joint Encoder
e.g., monoBERT

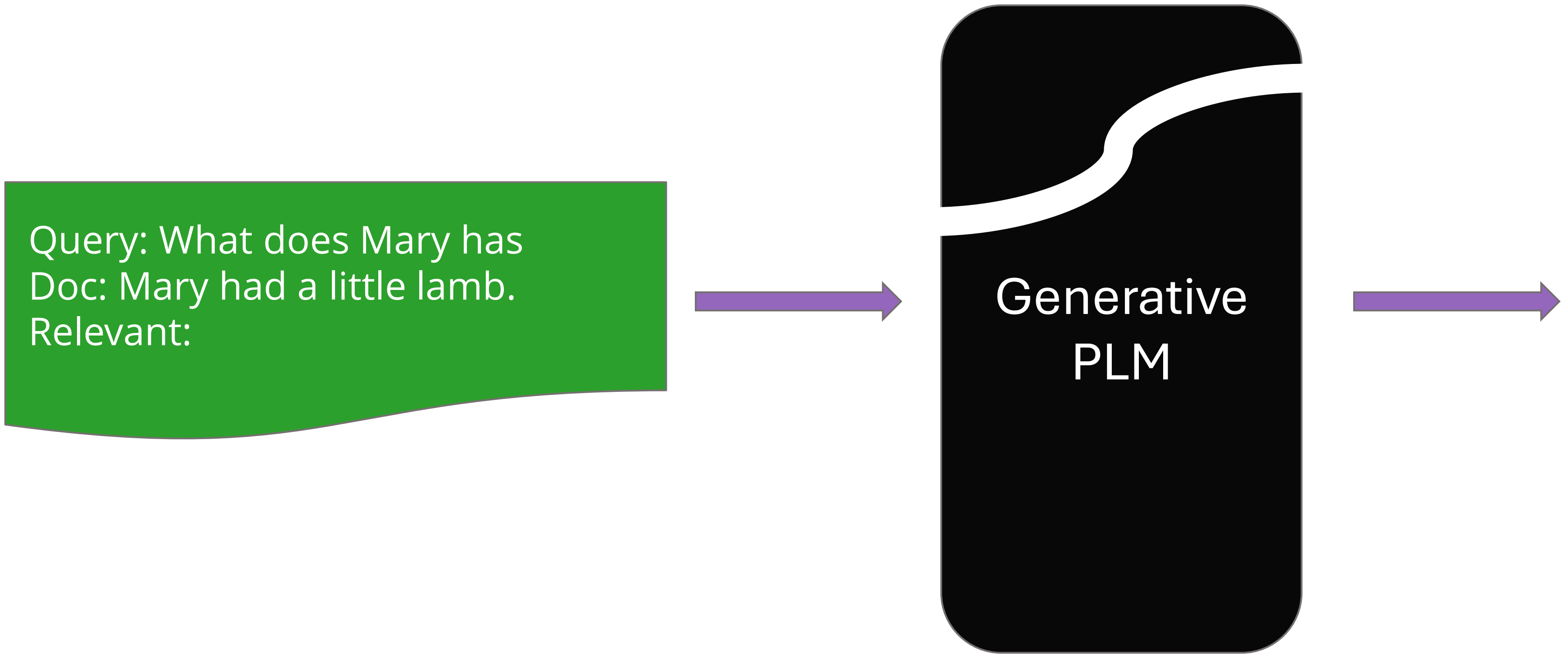
Cross-Encoder





Using Generative Models





Query: What does Mary has
Doc: Mary had a little lamb.
Relevant:

Generative
PLM

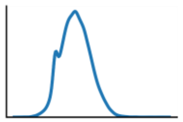
e.g, T5

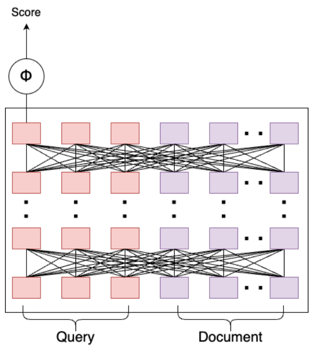
Not a number!

Yes

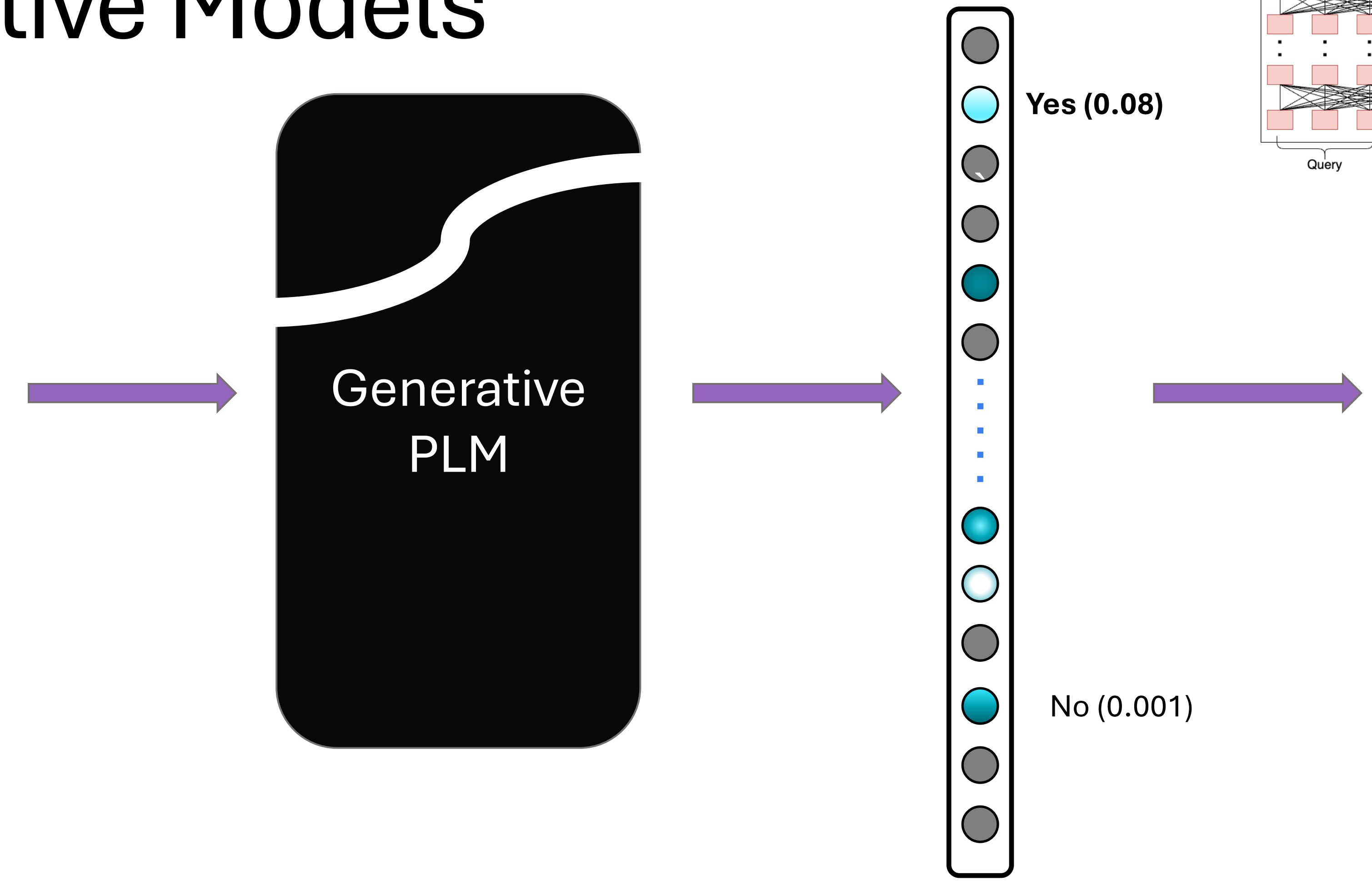
Pradeep, Ronak, Rodrigo Nogueira, and Jimmy Lin. "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models." arXiv preprint arXiv:2101.05667 (2021).

Using Generative Models





Query: What does Mary has
Doc: Mary had a little lamb.
Relevant:



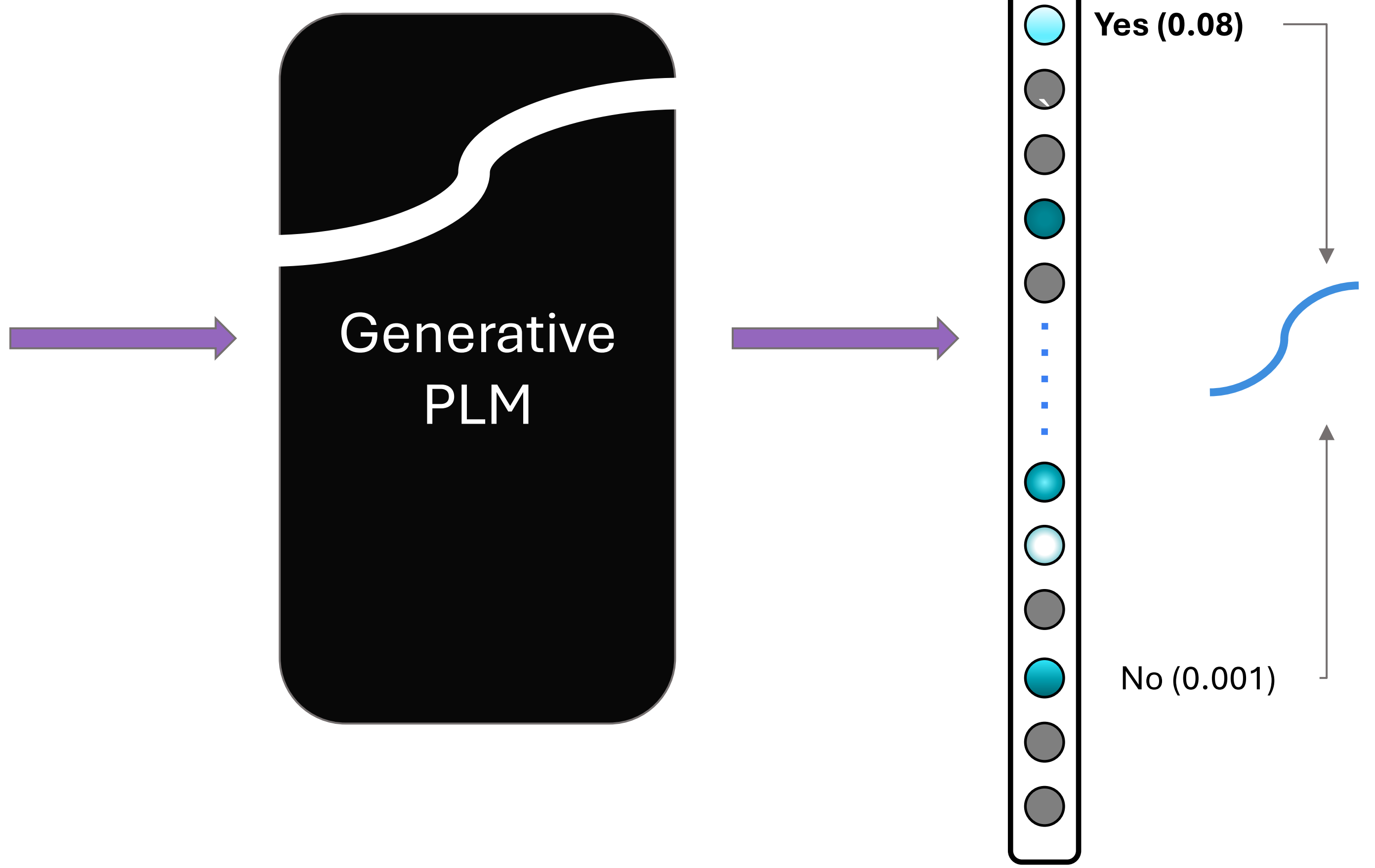
Yes

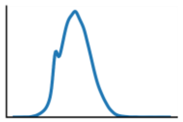
Pradeep, Ronak, Rodrigo Nogueira, and Jimmy Lin. "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models." arXiv preprint arXiv:2101.05667 (2021).

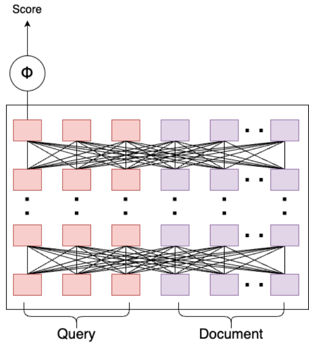
Using Generative Models

Pointwise score

Query: What does Mary has
Doc: Mary had a little lamb.
Relevant:





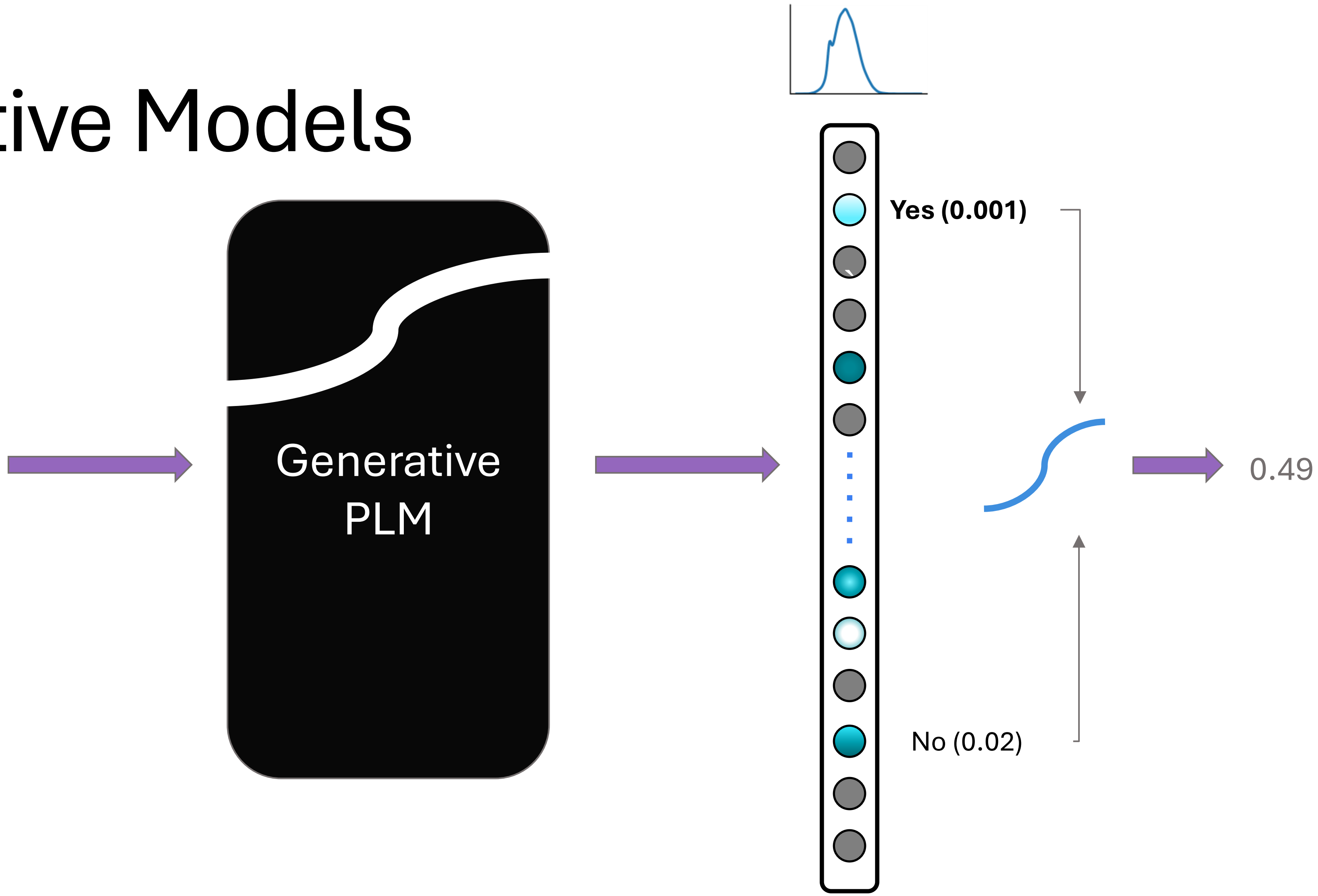


Pradeep, Ronak, Rodrigo Nogueira, and Jimmy Lin. "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models." arXiv preprint arXiv:2101.05667 (2021).

Using Generative Models

Pairwise score

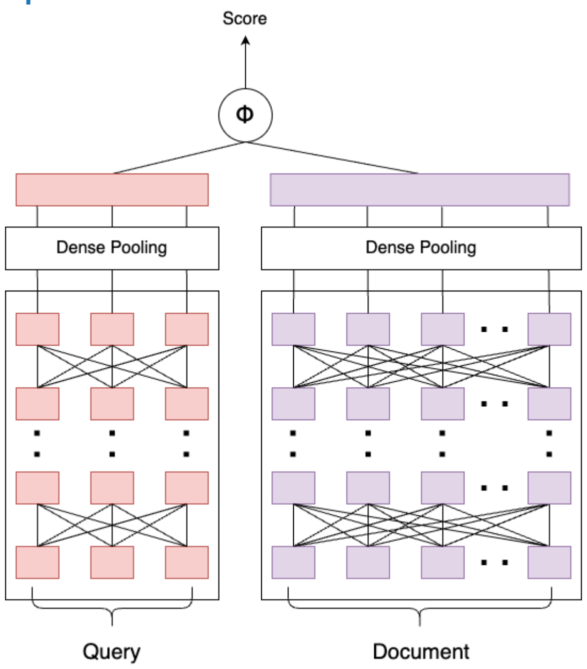
Query: What does Mary has
Doc0: JHU is in Baltimore
Doc1: Mary had a little lamb.
Relevant:



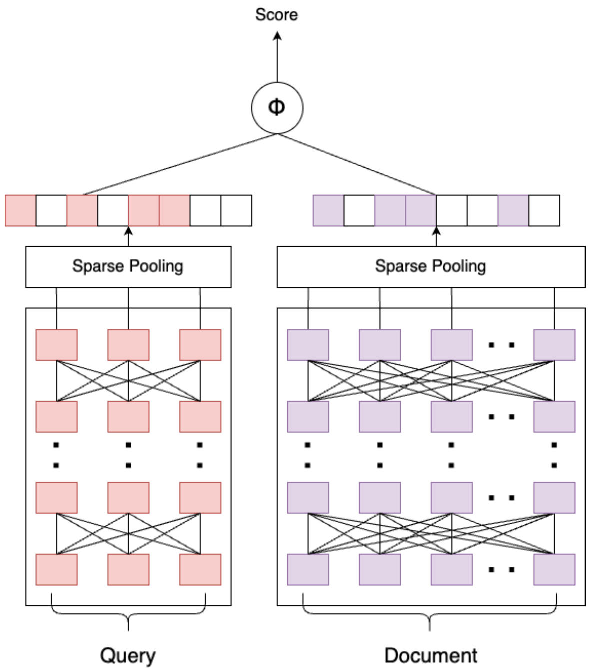
Pradeep, Ronak, Rodrigo Nogueira, and Jimmy Lin. "The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models." arXiv preprint arXiv:2101.05667 (2021).

Bi-Encoder

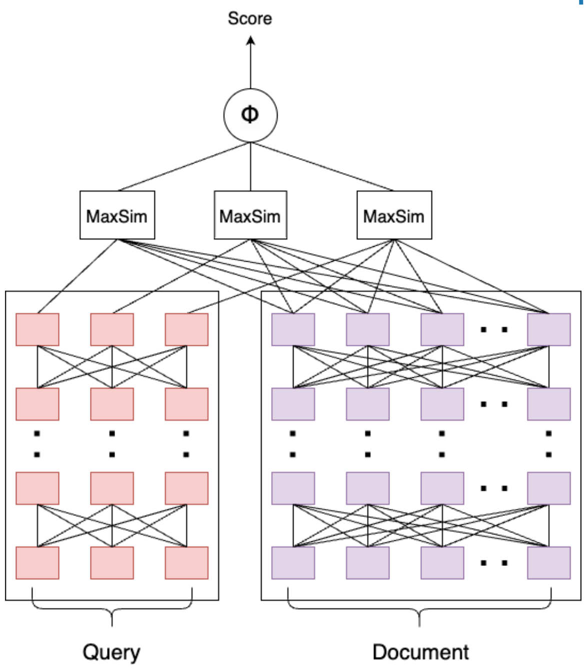
Cross Encoder



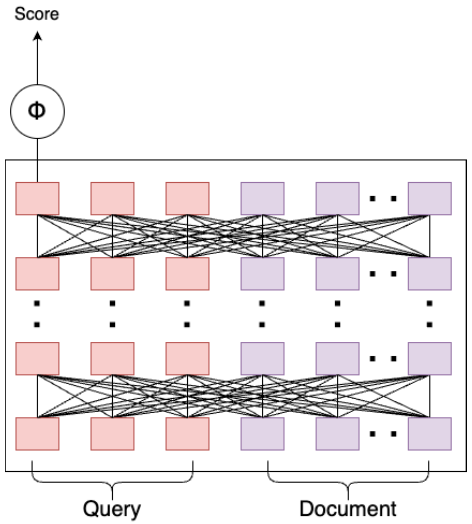
One Dense Vector
Per Sequence
e.g., DPR



One **Sparse** Vector
Per Sequence
e.g., SPLADE



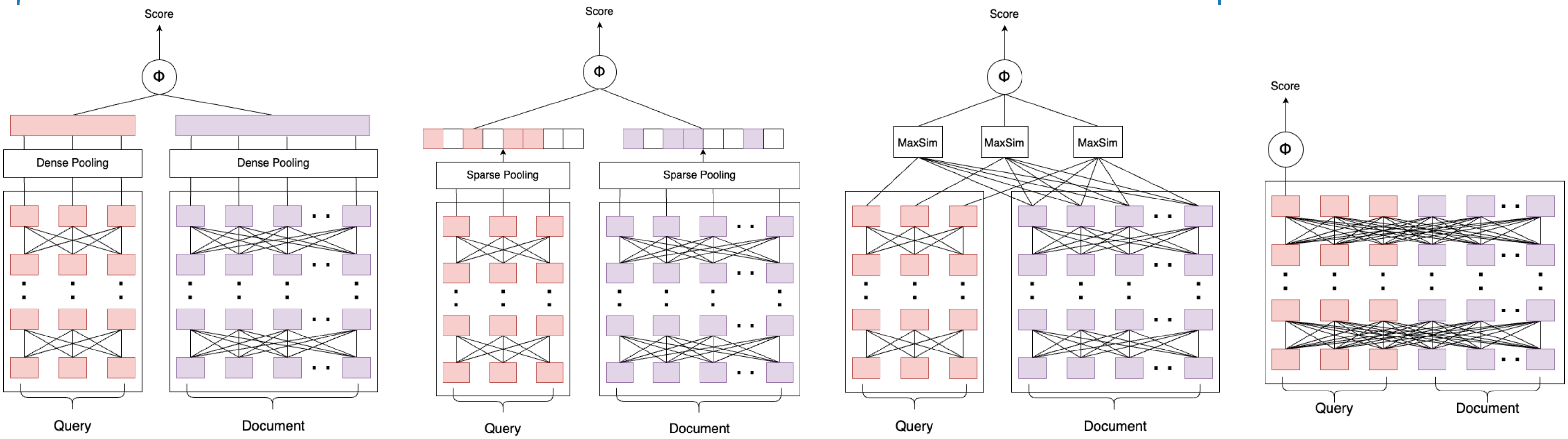
Multiple Dense Vectors
Per Sequence
e.g., ColBERT



Joint Encoder
e.g., monoBERT

Bi-Encoder

Cross Encoder



One Dense Vector
Per Sequence
e.g., DPR

One **Sparse** Vector
Per Sequence
e.g., SPLADE

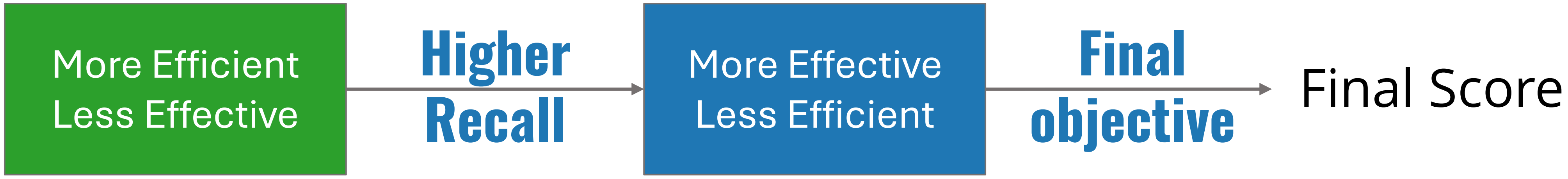
Multiple Dense Vectors
Per Sequence
e.g., ColBERT

Joint Encoder
e.g., monoBERT

More Effective

More Efficient

Retrieve-and-Rerank System Combinations



More Efficient
Less Effective

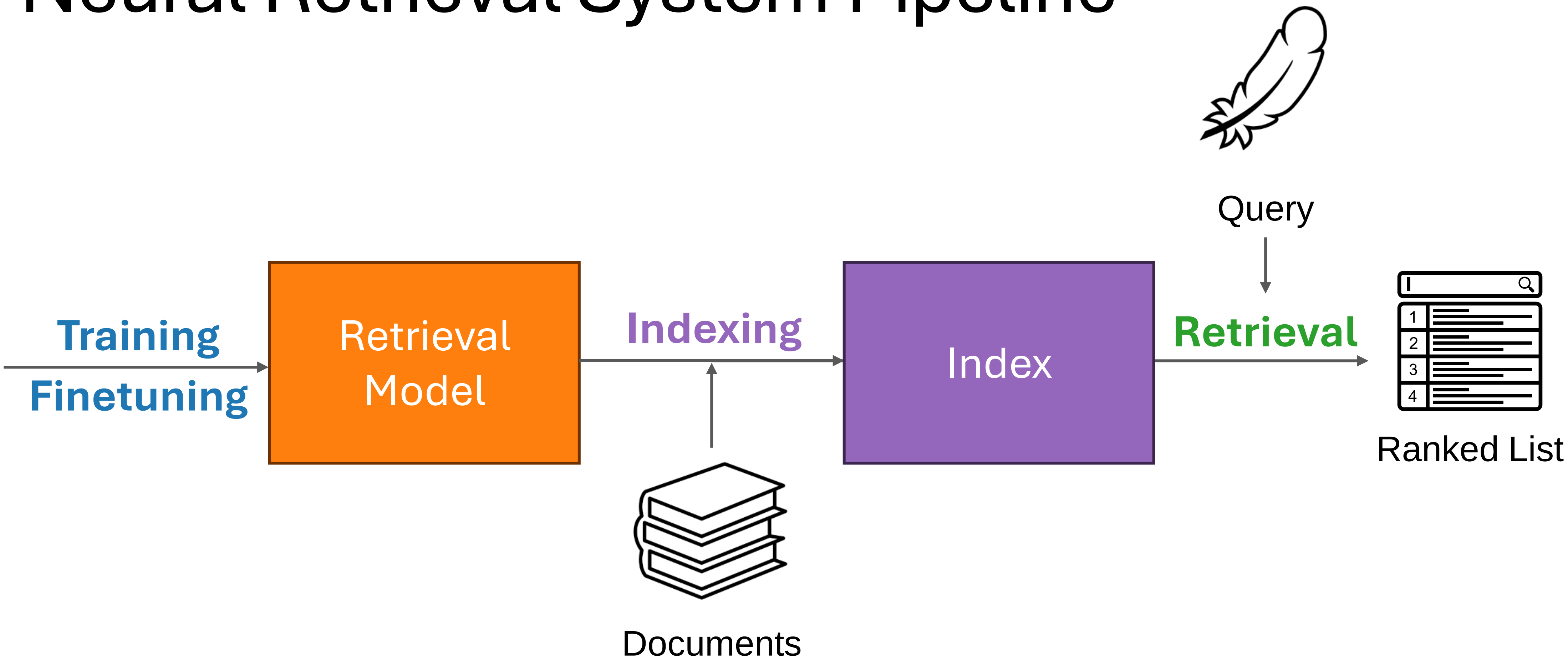
Higher Recall

More Effective
Less Efficient

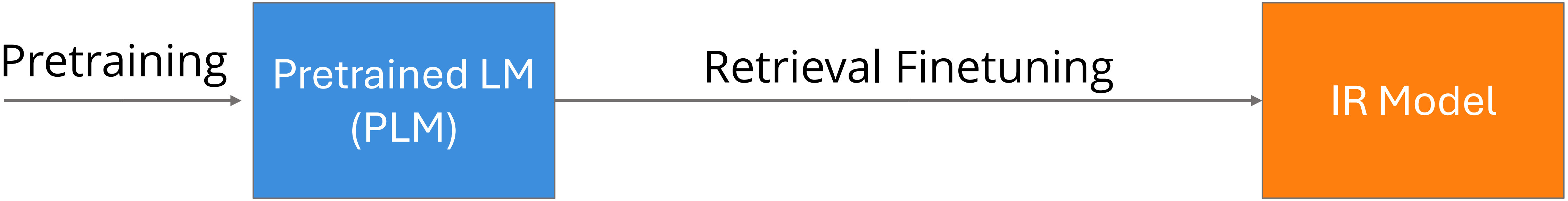
Final objective

Final Score

Neural Retrieval System Pipeline



PLM to IR Model



- Align the representation
- Model “relevancy”

Evaluation

Which system is better?

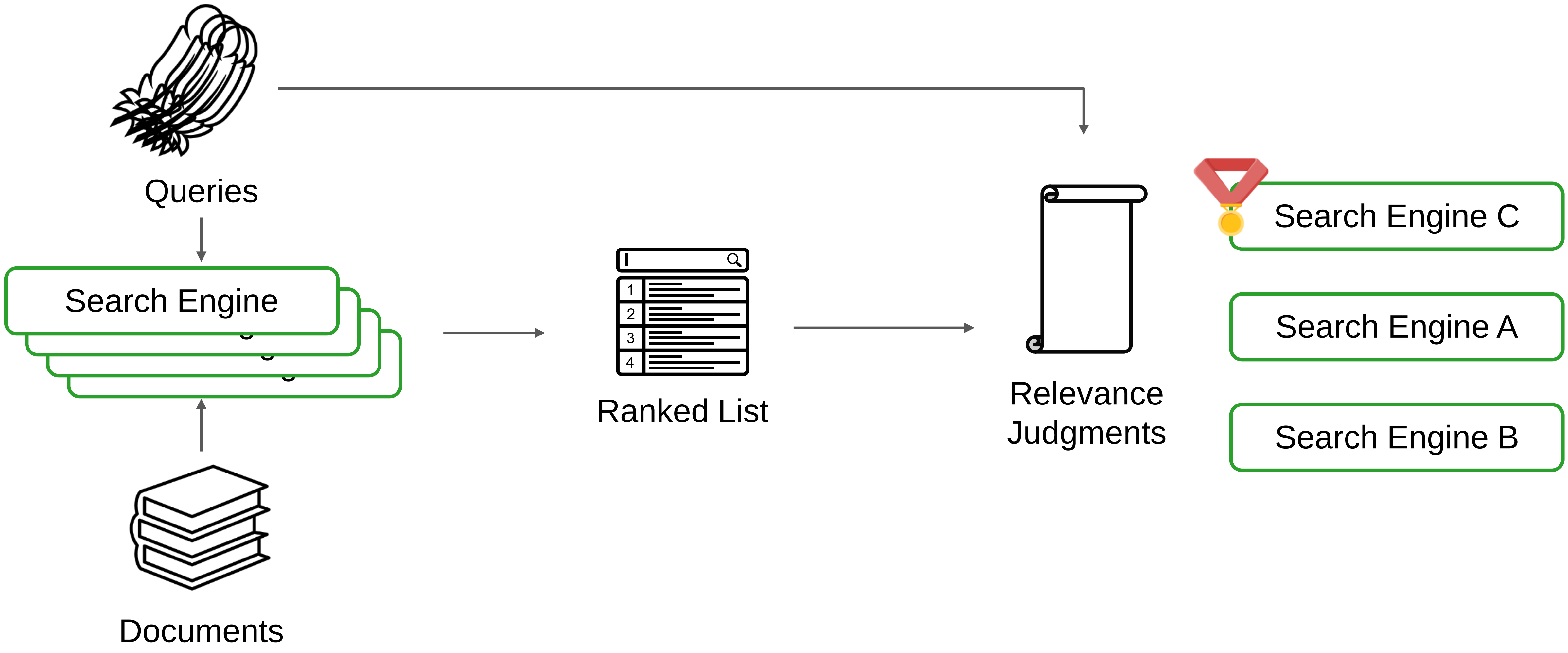
What is Information Retrieval?

(relevant)

Retrieve information from a storage based on user's information need

Which system retrieve more relevant information?

Cranfield Paradigm Evaluation



Queries

Search Engine

Documents

Ranked List

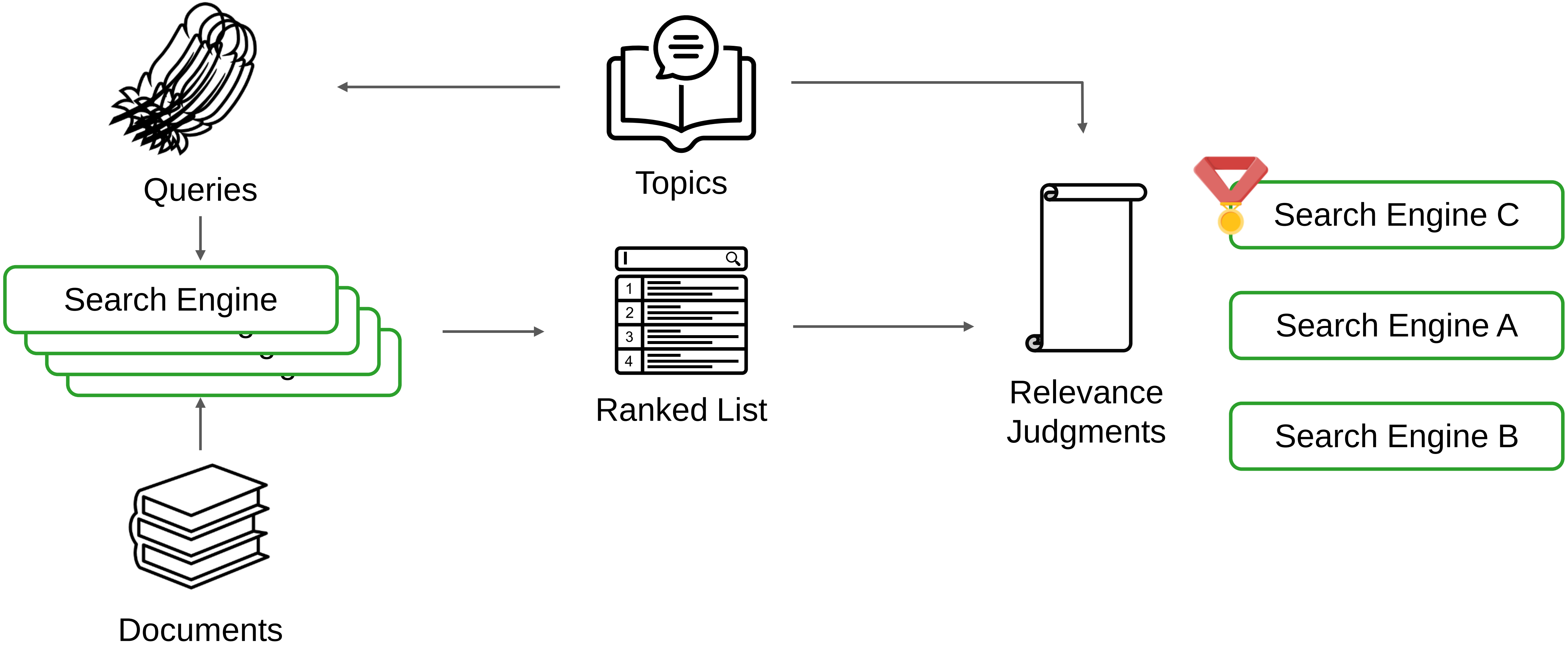
Relevance Judgments

Search Engine C

Search Engine A

Search Engine B

Cranfield Paradigm Evaluation



Queries

Topics

Search Engine

Ranked List

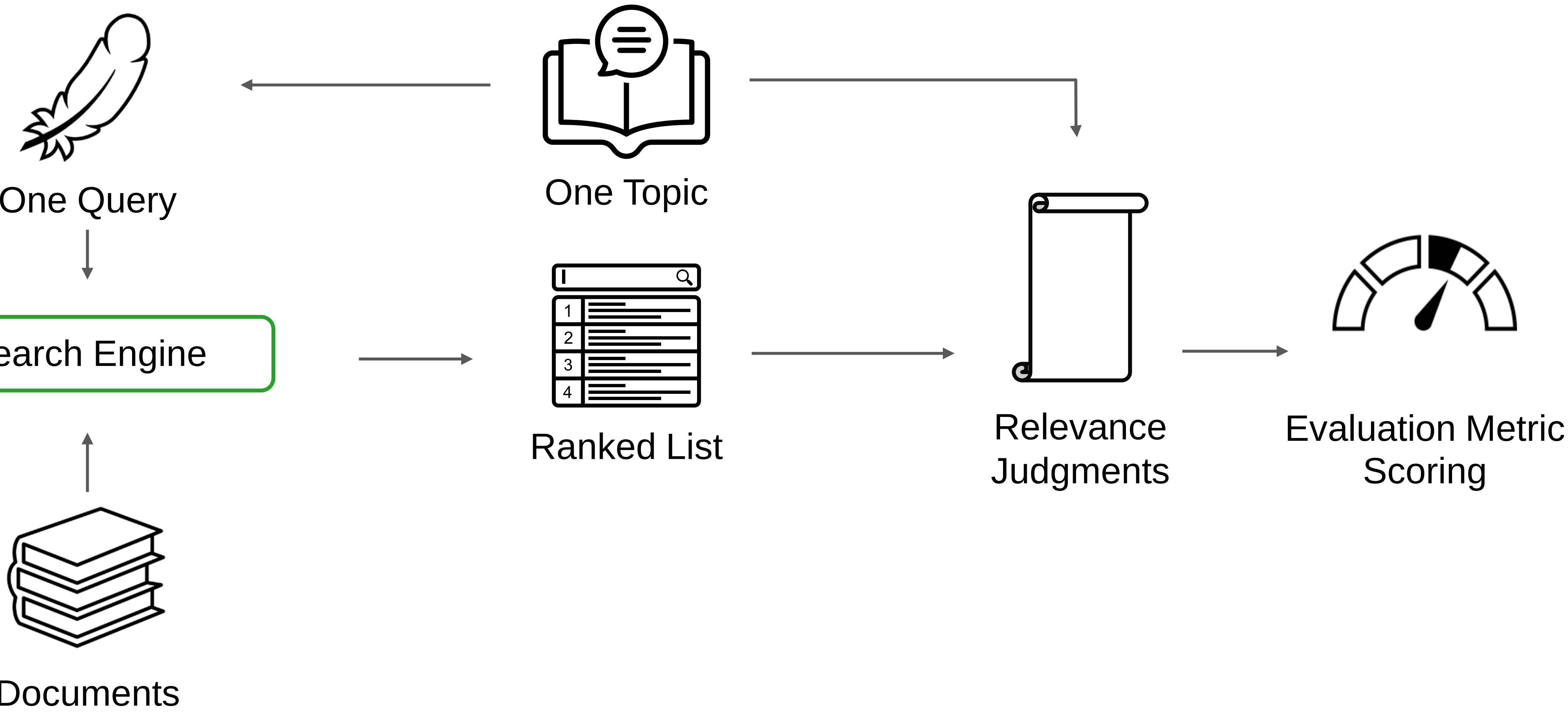
Relevance Judgments

Search Engine C

Search Engine A

Search Engine B

Cranfield Paradigm Evaluation



One Query

One Topic

Search Engine

Ranked List

Relevance Judgments

Evaluation Metric Scoring

Documents

IR-Specific Issues

- Topics vs Queries
  - Clear intent vs an expression of such intent
- Relevant vs related
  - Fulfilling the information need or not
- Relevance Judgements vs Labels
  - Opinion vs “fact”
- Ranked retrieval metrics
  - Measuring the quality/effectiveness of a ranked list

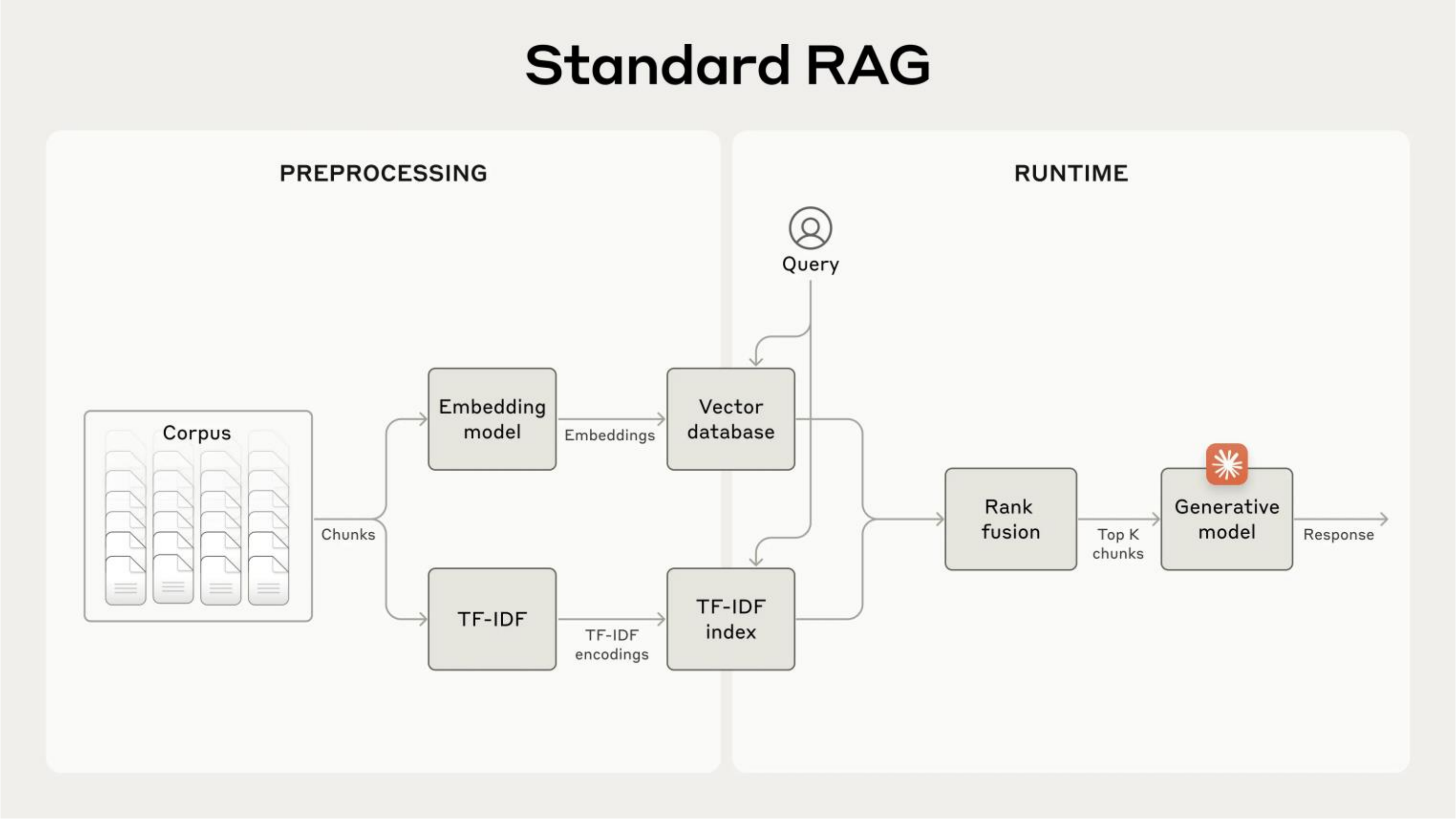
IR Metrics

- Effective Metrics
    - Mean Average Precision
    - Normalized Discounted Cumulative Gain
    - Recall@k
- Efficiency Metrics
    - Indexing time
    - Index disk space
    - Query latency (average search time per query)

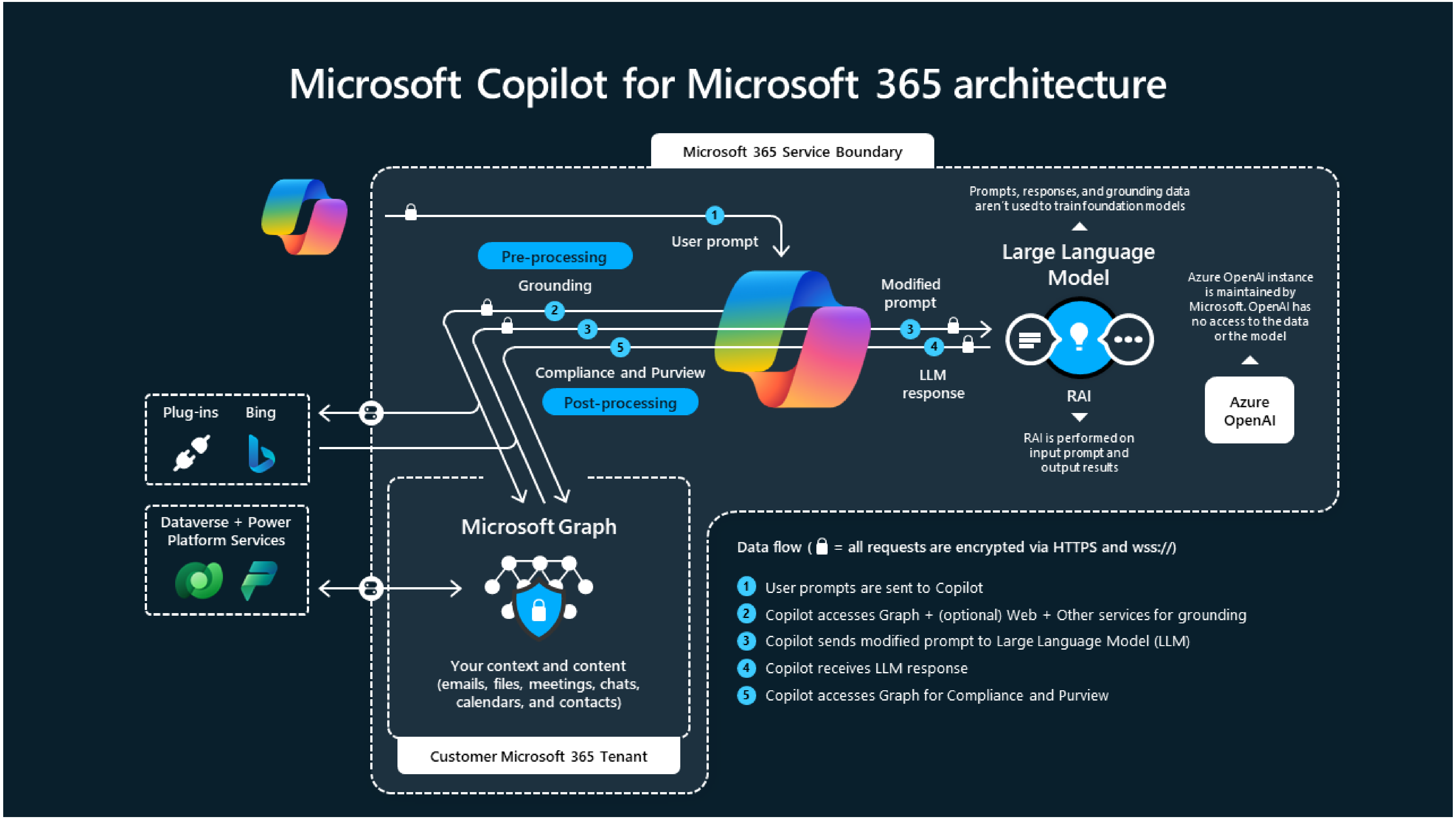
State of IR Research

Retrieval-Augmented Generation

- Is everything a RAG problem?
- What is the right retrieval model/system for RAG?
- IR going away?



https://www.anthropic.com/news/contextual-retrieval



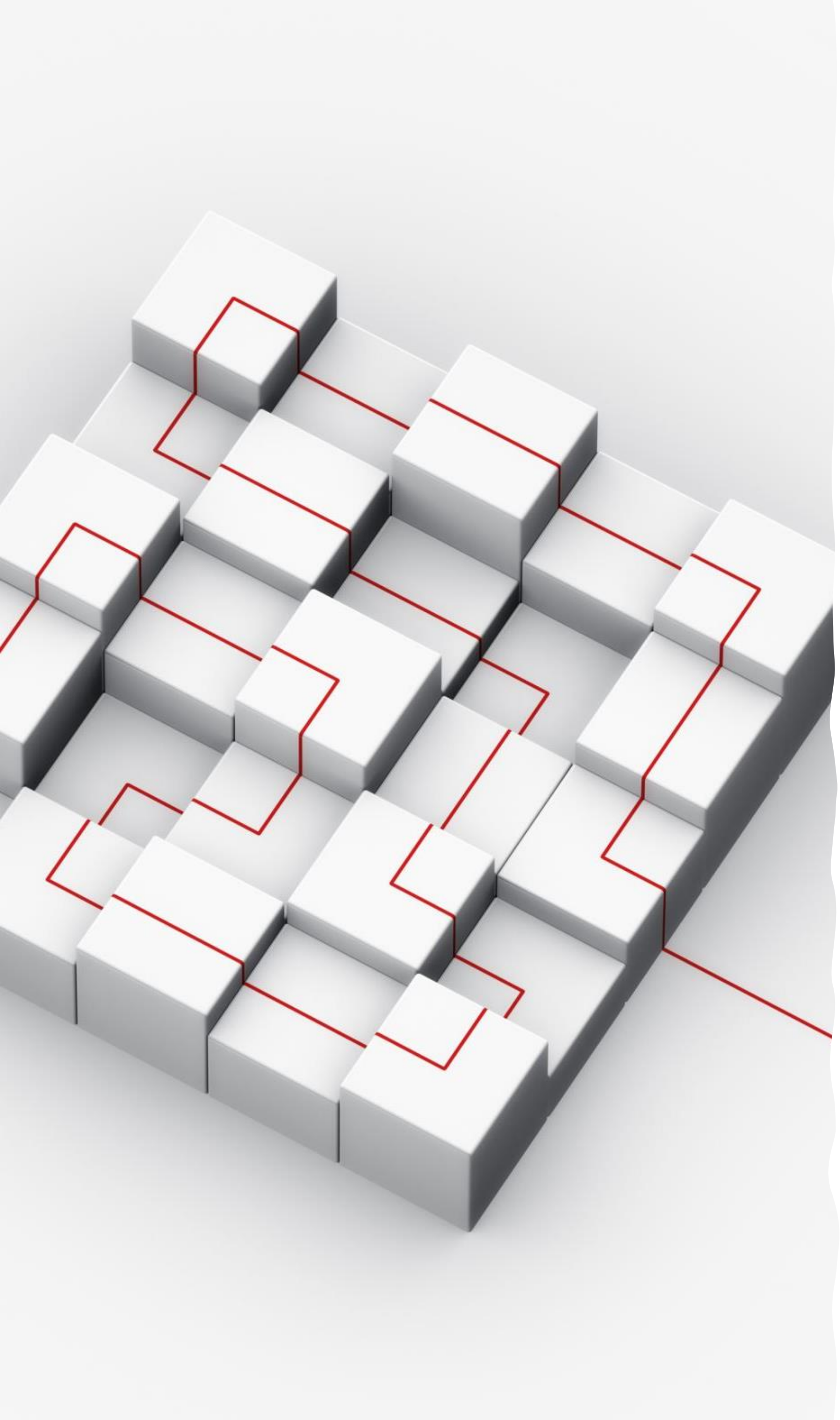
Microsoft Copilot for Microsoft 365 architecture

Microsoft 365 Service Boundary
User prompt
Pre-processing
Grounding
Compliance and Purview
Post-processing
Modified prompt
LLM response
Prompts, responses, and grounding data aren't used to train foundation models
Large Language Model
RAI
RAI is performed on input prompt and output results
Azure OpenAI instance is maintained by Microsoft. OpenAI has no access to the data or the model
Azure OpenAI
Plug-ins
Bing
Dataverse + Power Platform Services
Microsoft Graph
Your context and content (emails, files, meetings, chats, calendars, and contacts)
Customer Microsoft 365 Tenant
1
2
3
4
5

Data flow (🔒 = all requests are encrypted via HTTPS and wss://)

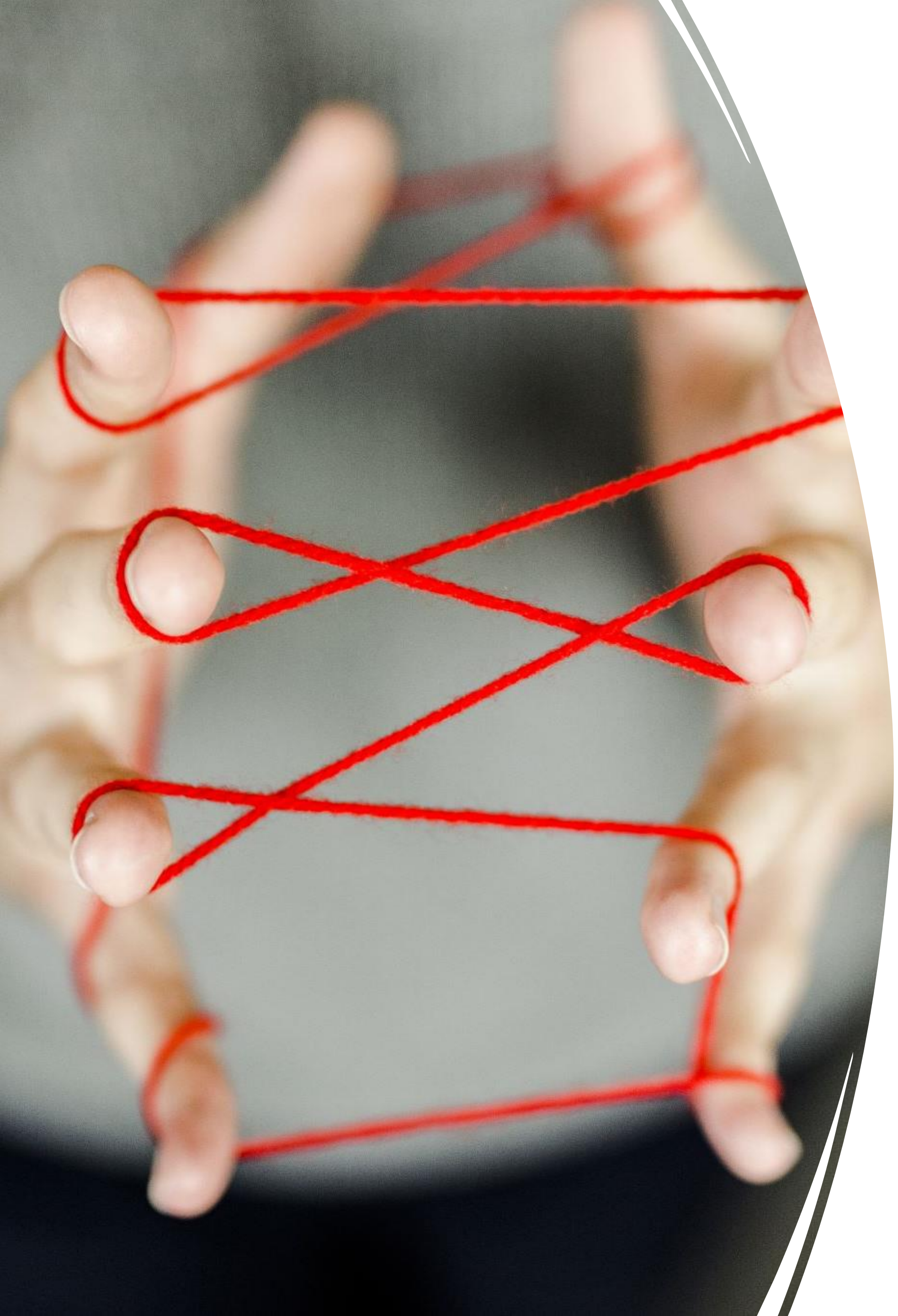
1. User prompts are sent to Copilot
2. Copilot accesses Graph + (optional) Web + Other services for grounding
3. Copilot sends modified prompt to Large Language Model (LLM)
4. Copilot receives LLM response
5. Copilot accesses Graph for Compliance and Purview

https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-overview



Better Retrieval Models

- More effective
  - Better/larger neural models
  - Better architecture?
  - Under harder setup, e.g., scholar search, multilingual, cross-modal, etc
- More efficient
  - Faster at query time
  - Less resource footprint, e.g., memory, storage, compute, etc
- Other qualities
  - Fairness, diversity, etc



Other Retrieval Problems

- Conversational
  - Guessing intent, finding the "right" information to serve
- Iterative/interactive/human-in-the-loop
  - Rounds of interactions
- Generative
  - Returning a piece of text

Evaluation

- What to measure
  - and when would it fail
- How to measure
  - Generative text? Citations?
- “Better” evaluation collection
  - Not necessarily larger