Retrieval Augmented Generation

Liad Magen

Vienna Deep-Learning Meetup

April 2024

Surfaces and Interfaces
Volume 46, March 2024, 104081

The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries

Manshu Zhang [a 1], Liming Wu [a 1], Tao Yang [b], Bing Zhu [a], Yangai Liu [a]

Show more

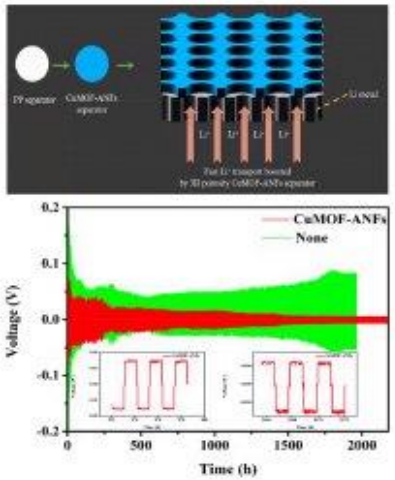
https://doi.org/10.1016/j.surfin.2024.104081

Abstract

Lithium metal, due to its advantages of high theoretical capacity, low density and low electrochemical reaction potential, is used as a negative electrode material for batteries

density. The results show that CuMOF-ANFs composite membrane can inhibit the generation of lithium dendrites and improve the cycle stability and cycle life of the battery. The three-dimensional (3D) porous mesh structure of CuMOF-ANFs separator provides a new perspective for the practical application of lithium metal battery.

Graphical abstract



Download : Download high-res image (180KB)

Download : Download full-size image

Introduction

Certainly, here is a possible introduction for your topic:Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high theoretical capacities [1], [2]. However, during the cycle, dendrites forming on the lithium metal anode can cause a short circuit, which can affect

The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries - ScienceDirect

ELSEVIER

Surfaces and Interfaces

Volume 46, March 2024, 104081

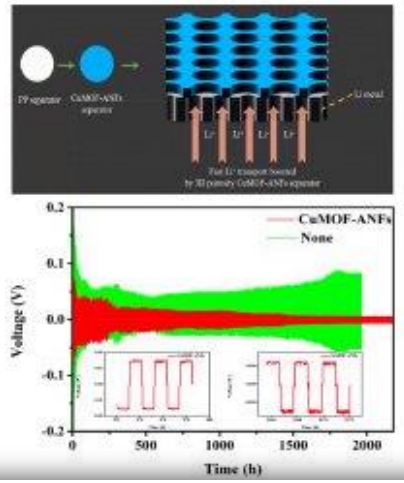
The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries

Manshu Zhang [a 1], Liming Wu [a 1], Tao Yang [b], Bing Zhu [a], Yangai Liu [a]

Show more

density. The results show that CuMOF-ANFs composite membrane can inhibit the generation of lithium dendrites and improve the cycle stability and cycle life of the battery. The three-dimensional (3D) porous mesh structure of CuMOF-ANFs separator provides a new perspective for the practical application of lithium metal battery.

Graphical abstract



Introduction

Certainly, here is a possible introduction for your topic:Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low

separator enhances the electrochemical performance of lithium metal anode batteries - ScienceDirect

promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high theoretical capacities [1], [2]. However, during the cycle, dendrites forming on the lithium metal anode can cause a short circuit, which can affect

Everyone is using LLM Chat-Bots

Radiology Case Reports
Volume 19, Issue 6, June 2024, Pages 2106-2111

Case Report

Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review

Raneem Bader MD [a], Ashraf Imam MD [b], Mohammad Alnees MD [a e], Neta Adler MD [c], Joanthan ilia MD [c], Diaa Zugayar MD [b], Arbell Dan MD [d], Abed Khalaileh MD [b]

Show more

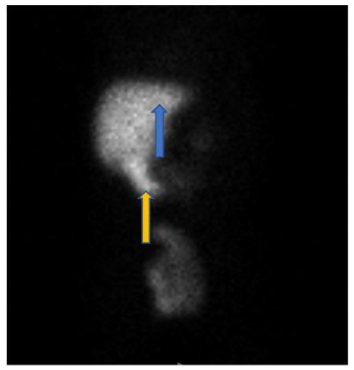
+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.radcr.2024.02.037

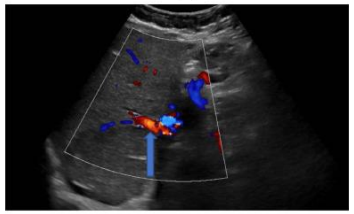
Get rights and content

Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review - ScienceDirect

The age of LLMs



(A)



(B)

prevention of complications such as cholangitis and biliary strictures.

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multidisciplinary team experienced in managing complex liver injuries.

Conclusion

In conclusion, proper treatment of iatrogenic vascular injuries is dependent on an

Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review - ScienceDirect

The age of LLMs

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multidisciplinary team experienced in managing complex liver injuries.

(B)

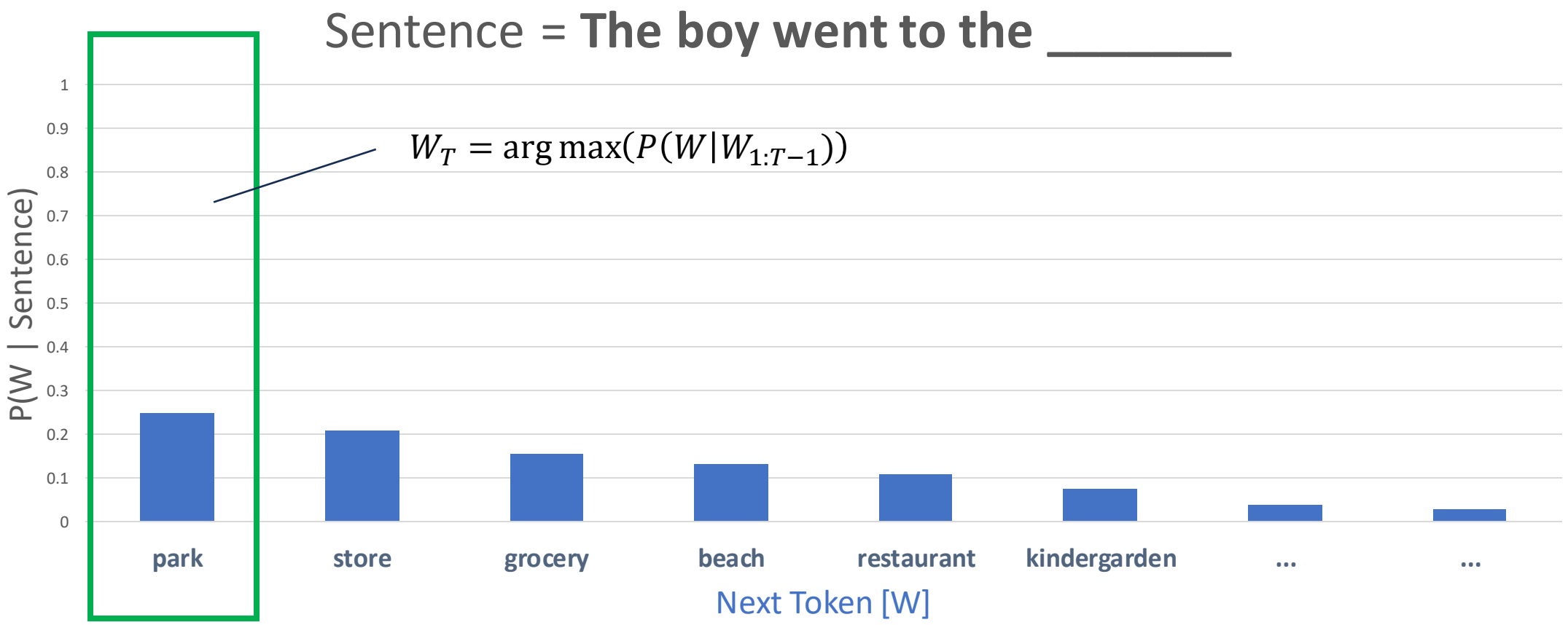
Conclusion

In conclusion, proper treatment of iatrogenic vascular injuries is dependent on an

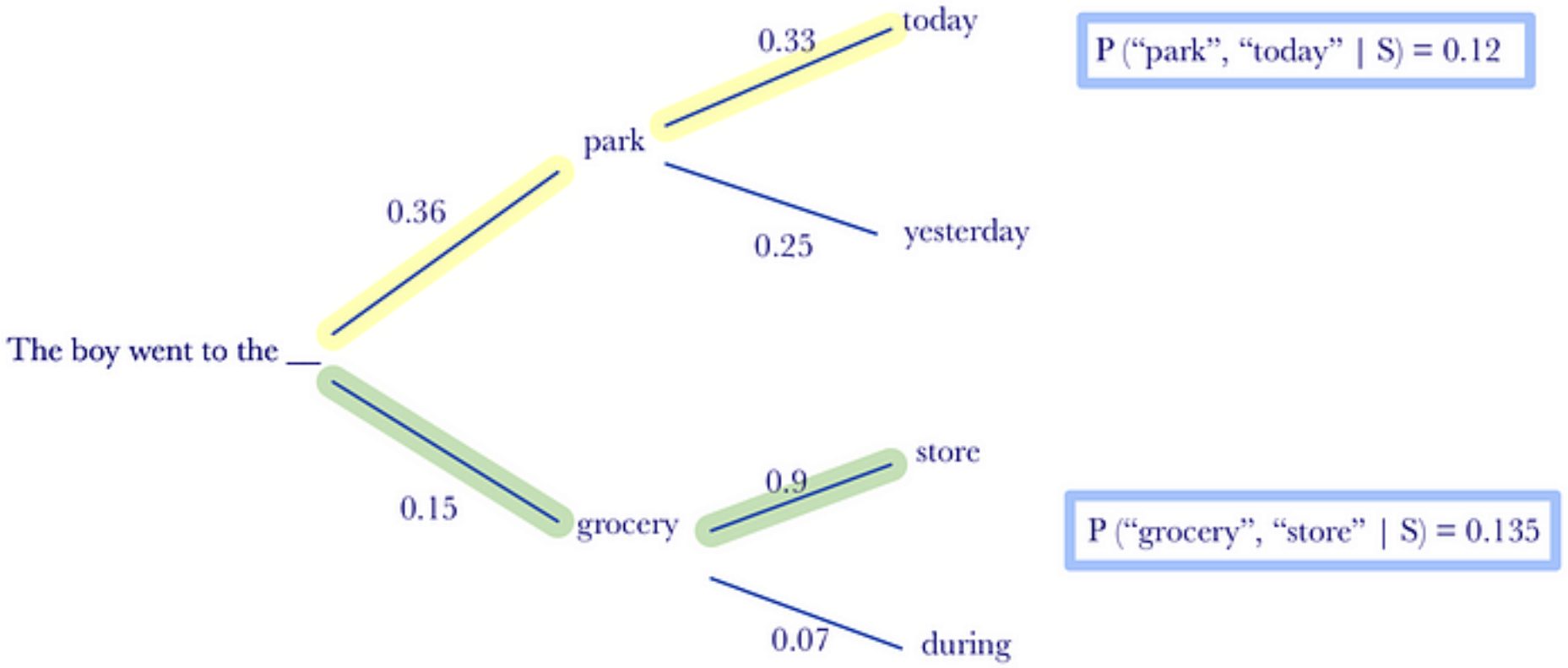
Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review - ScienceDirect

Access to Real-Time Information

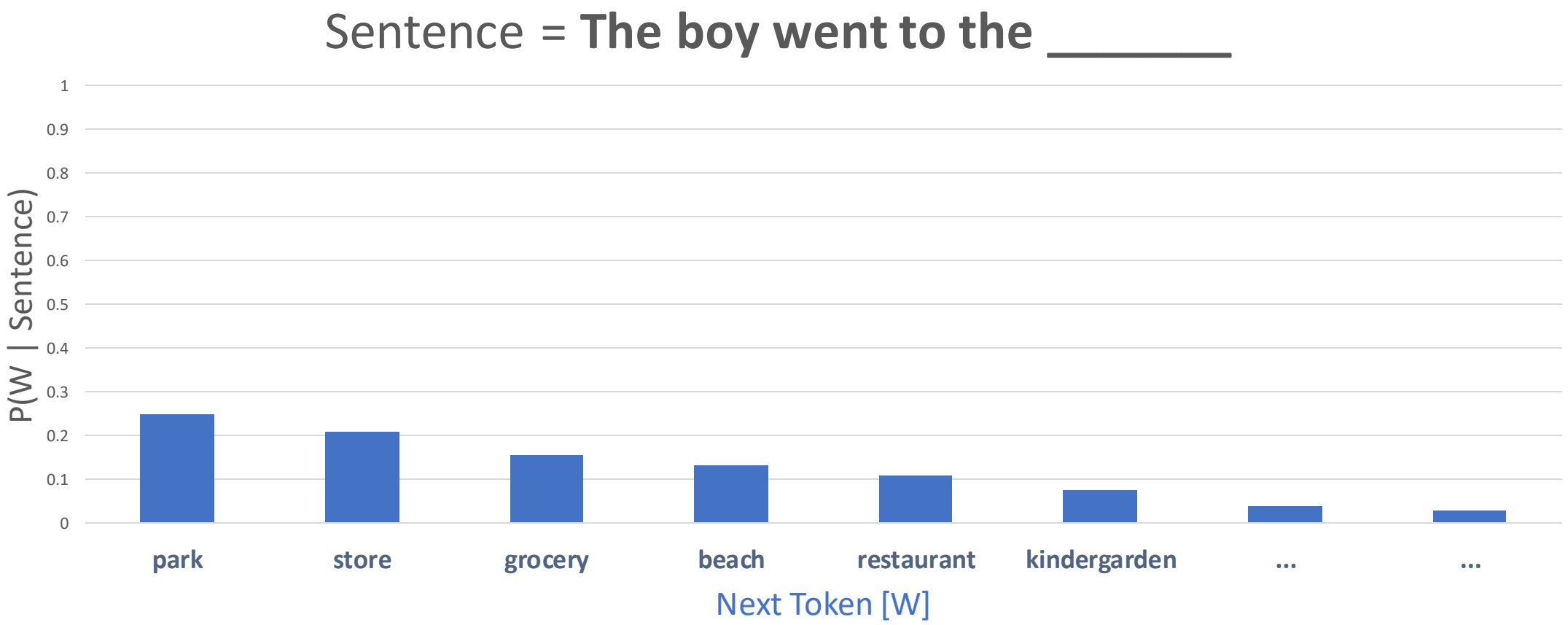
Language Models – A Short Reminder



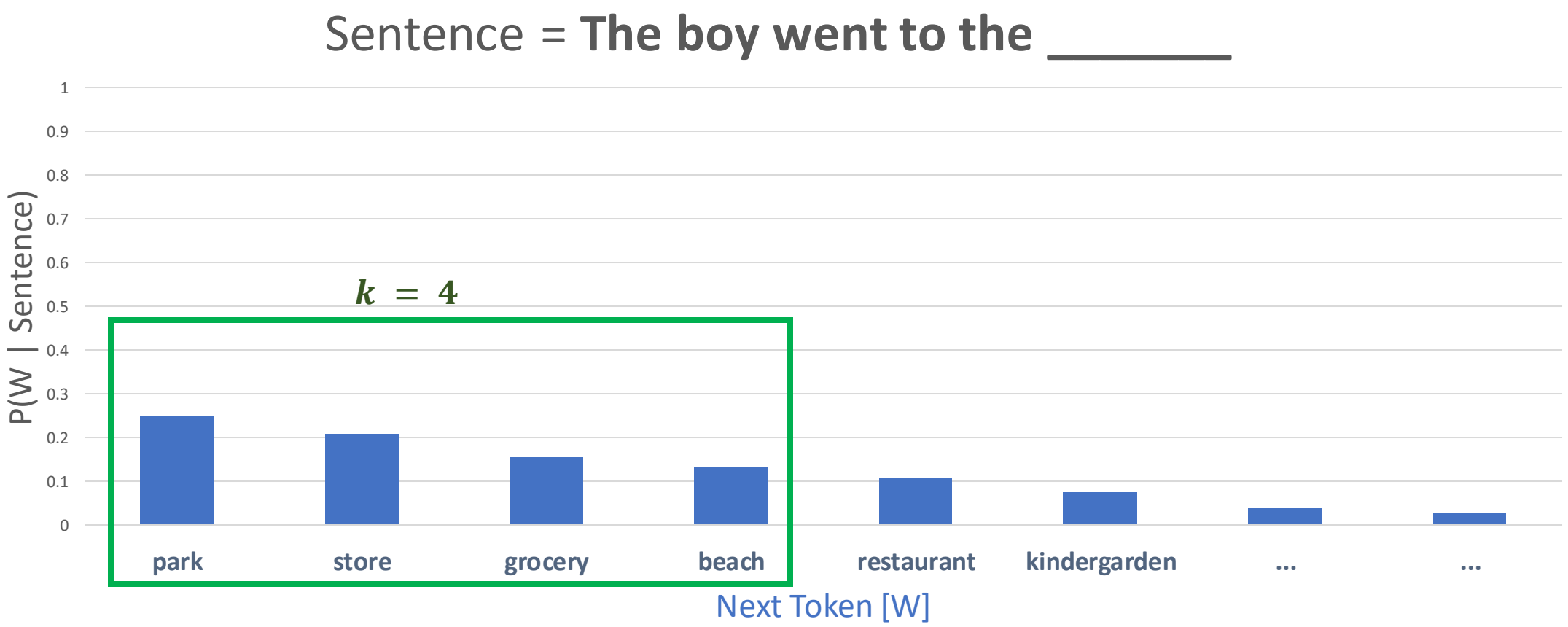
Language Models – A Short Reminder



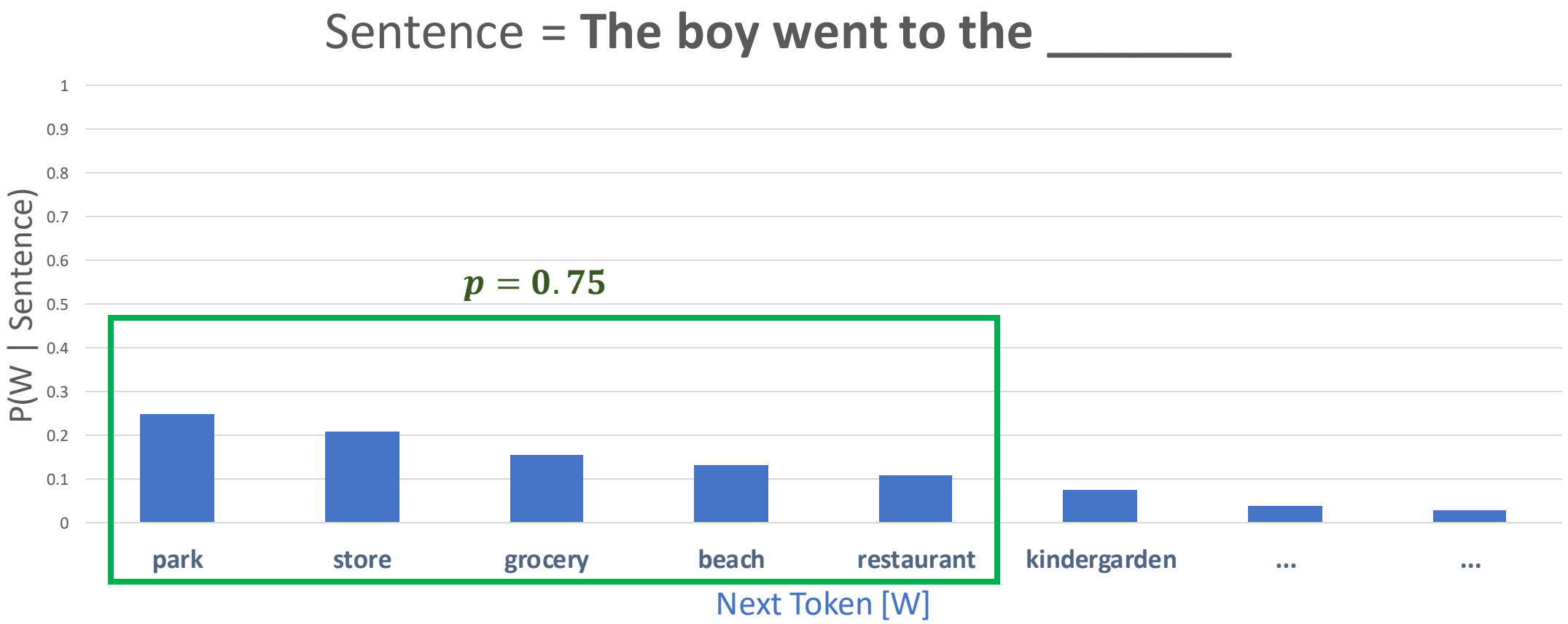
Language Models – Probability Distribution



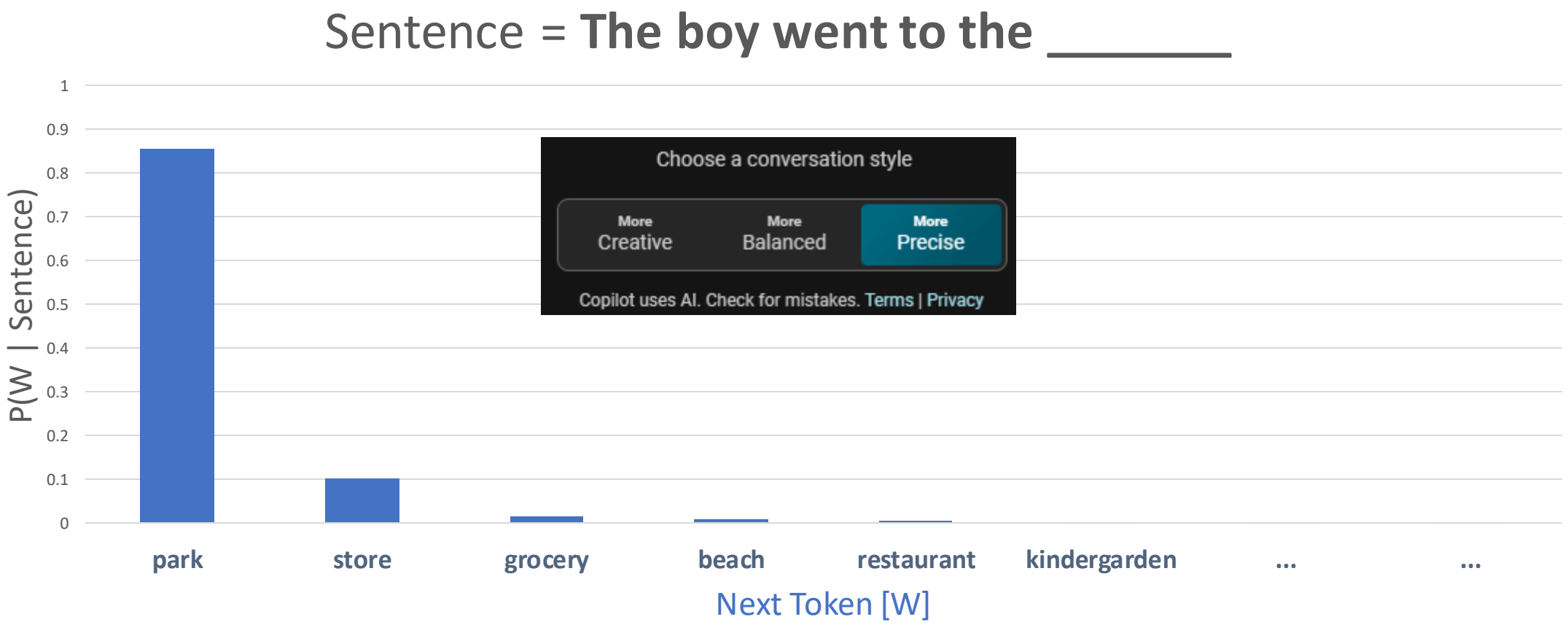
Language Models – Top K



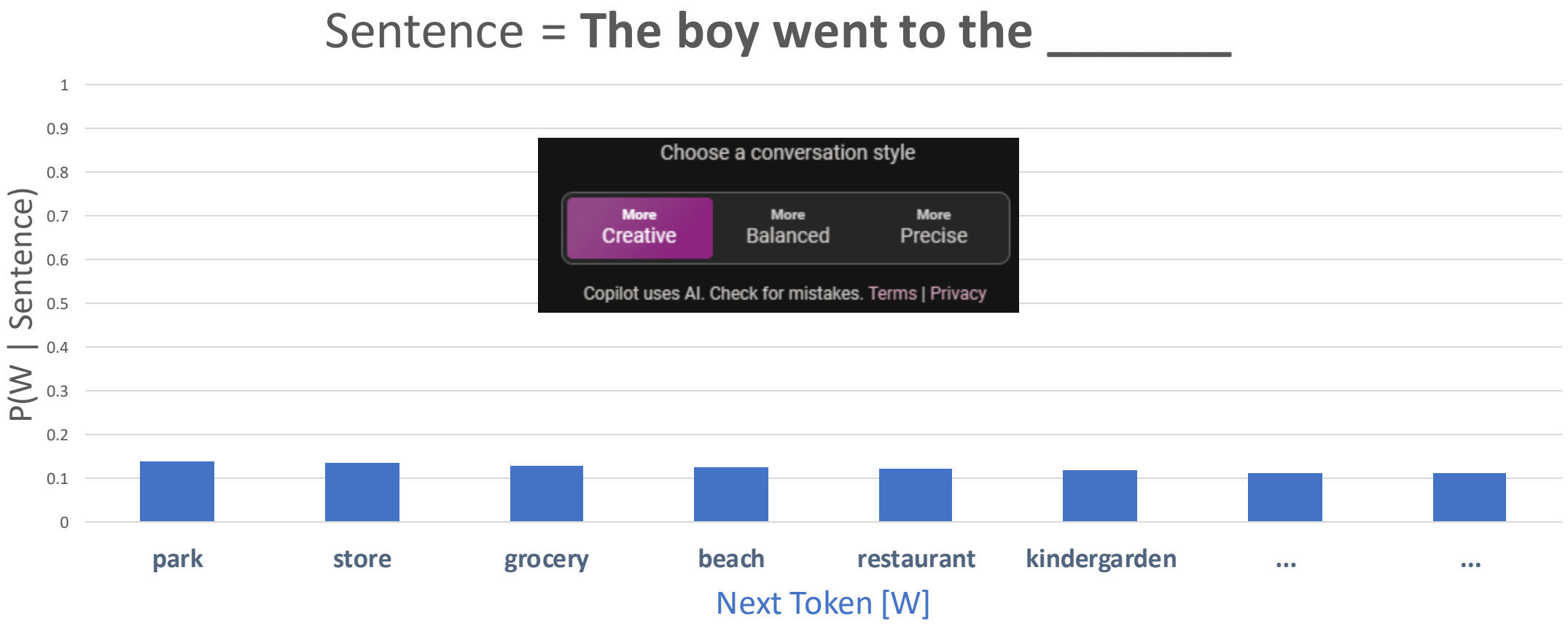
Language Models – Top P



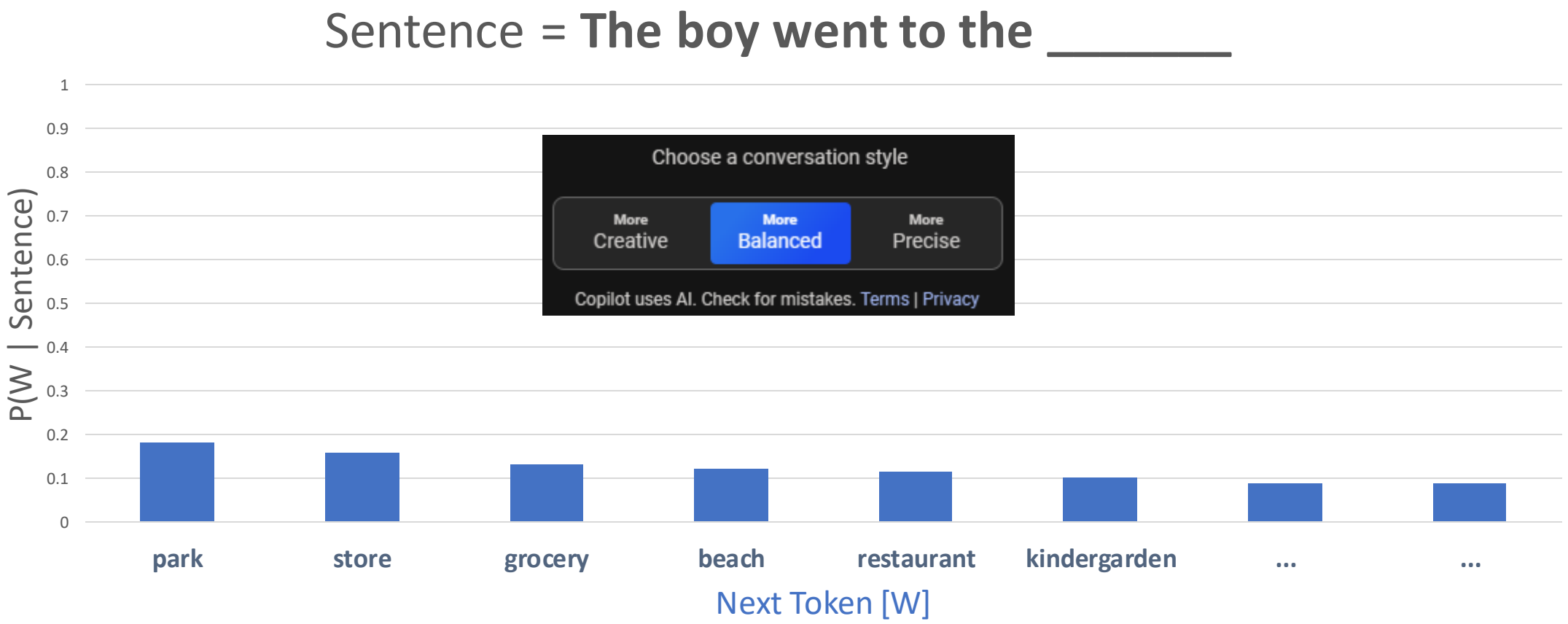
Language Models – Softmax Temperature (0.05)

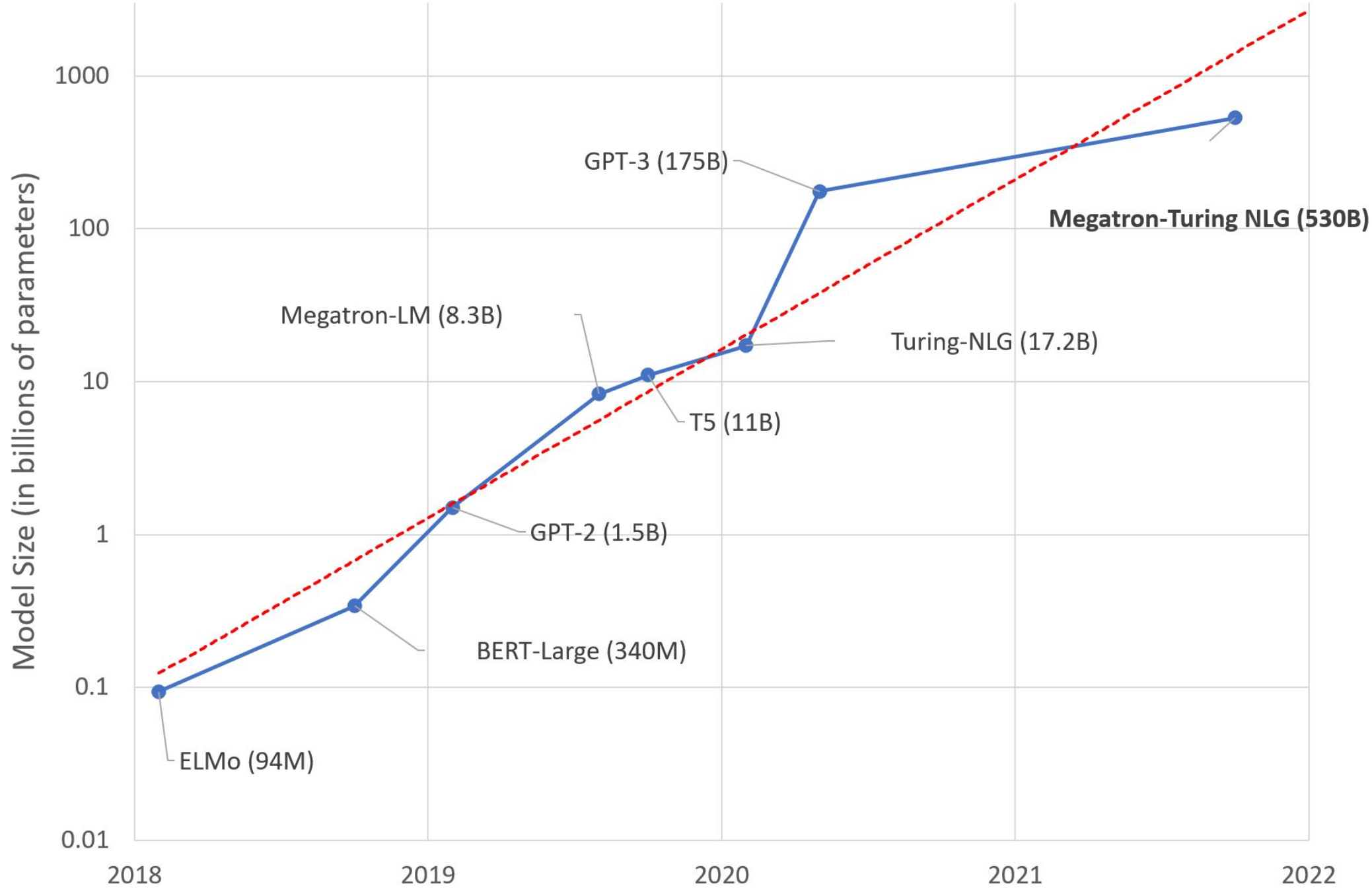


Language Models – Softmax Temperature (1.0)

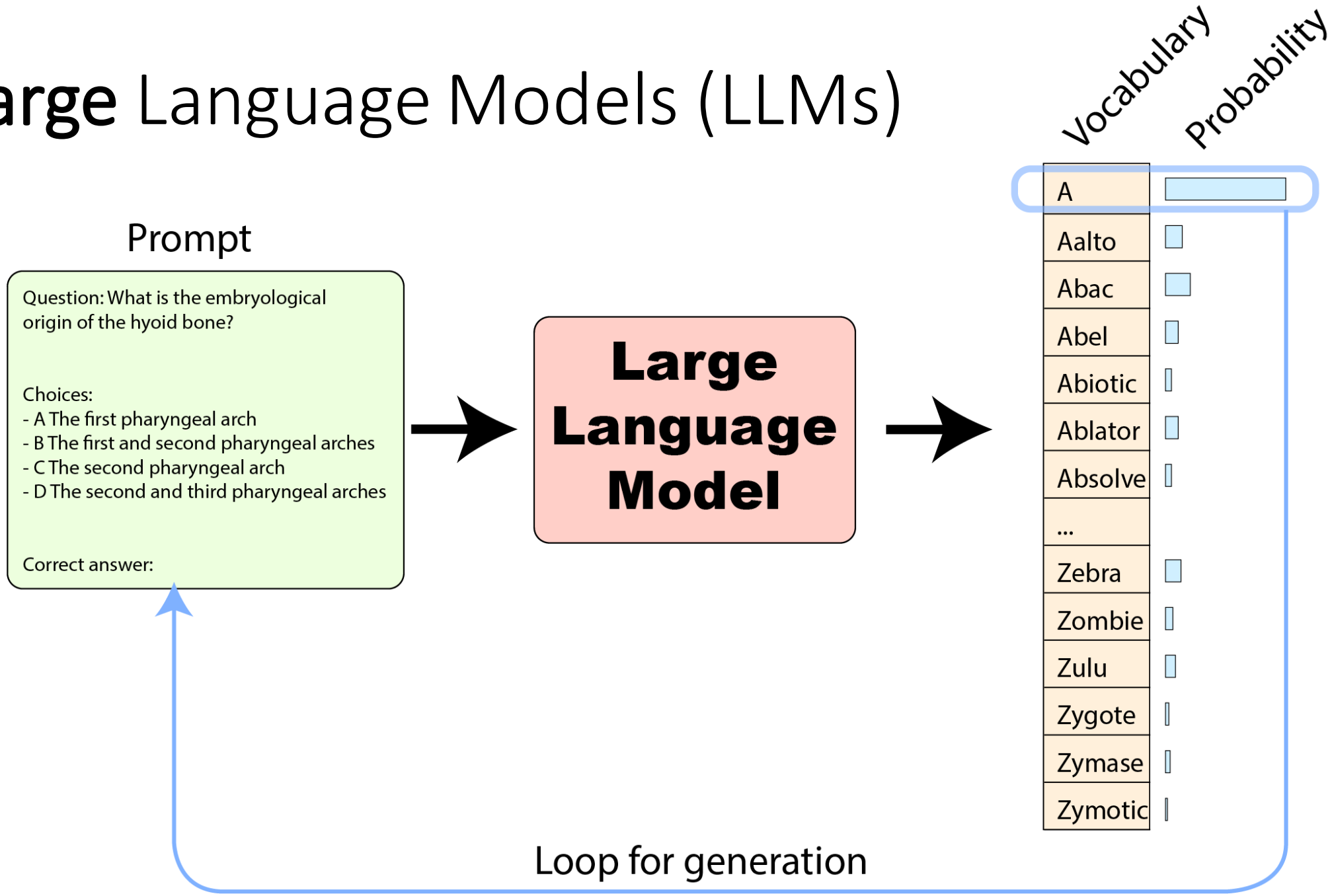


Language Models – Softmax Temperature (0.3)




Model Size (in billions of parameters)
1000
100
10
1
0.1
0.01
2018
2019
2020
2021
2022
GPT-3 (175B)
Megatron-Turing NLG (530B)
Megatron-LM (8.3B)
Turing-NLG (17.2B)
T5 (11B)
GPT-2 (1.5B)
BERT-Large (340M)
ELMo (94M)

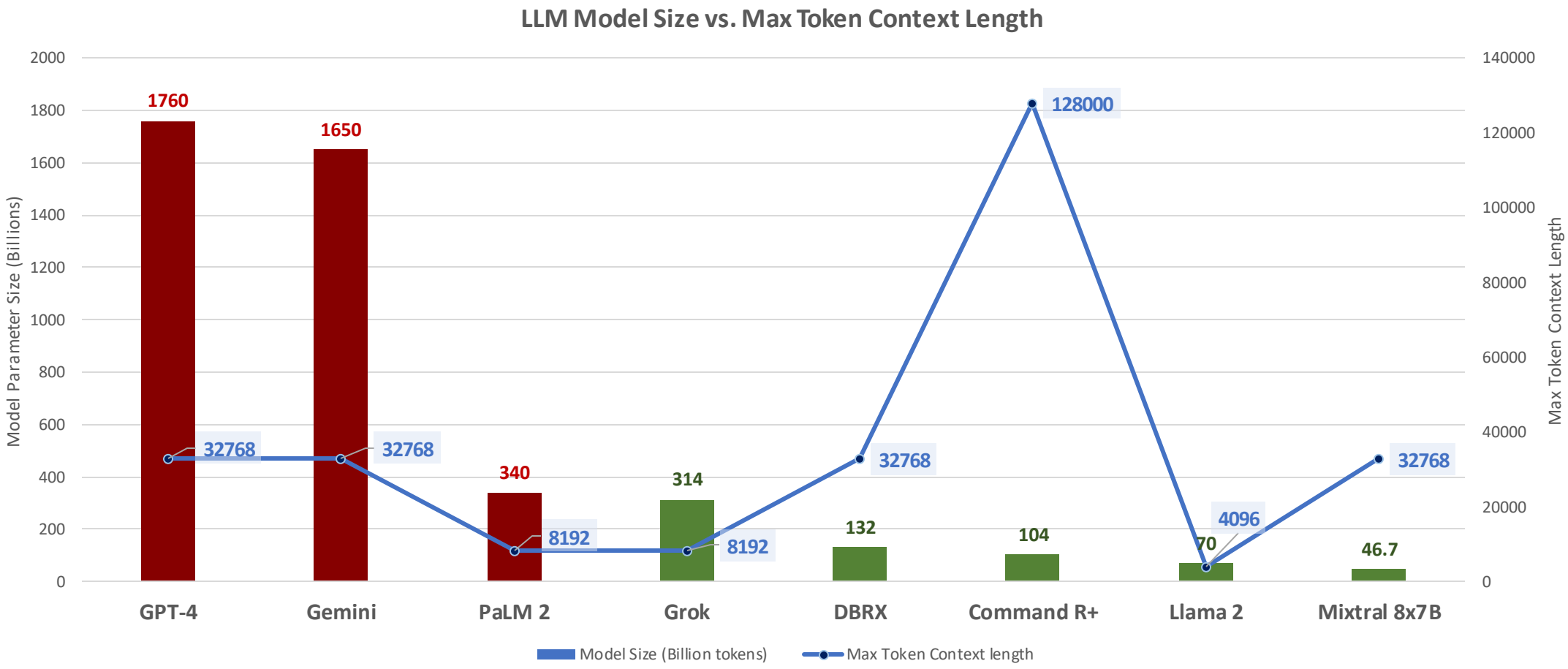
Large Language Models (LLMs)



LLMs - How Large?

	Llama 2	GPT-4	PaLM 2	Gemini	Mixtral 8x7B	Grok	DBRX	Command R+
Model Size	7 B 13 B 70 B	1,760 B	340 B	1,650 B 600 B 1.8 B	46.7B	314 B	132 B	104B
Max Token Context Length	4k	8k 32k	8k	32k	32k	8k	32k	128k
Modalities	Text	Text + Image	Text	Text + Image	Text	Text	Text	Text
	Open-Source	Closed	Closed	Closed	Open-Source	Open-Source	Open-Source	?

LLMs - How Large? - Visualization for Humans



What is your cut-off date?

My cut-off date is August 2023.

This means that my knowledge is based on information available up until that date.

I am unable to provide information or answer questions about events, developments, or changes that occurred after my cut-off date.

LE CHAT NOIR_

Drawbacks of LLMs

Limited proficiency in **specialized** domains

Outdated Information

Hallucination

Expensive to train

Hard to customize (fine-tune)

Lack of observability

Connect LLMs to an *external* data source

Why does it solve these problems?

Allows **customization**	• Without retraining
Provides **grounding**	• Citations • Less hallucinations
Prevents data leakage	• No training on sensitive data

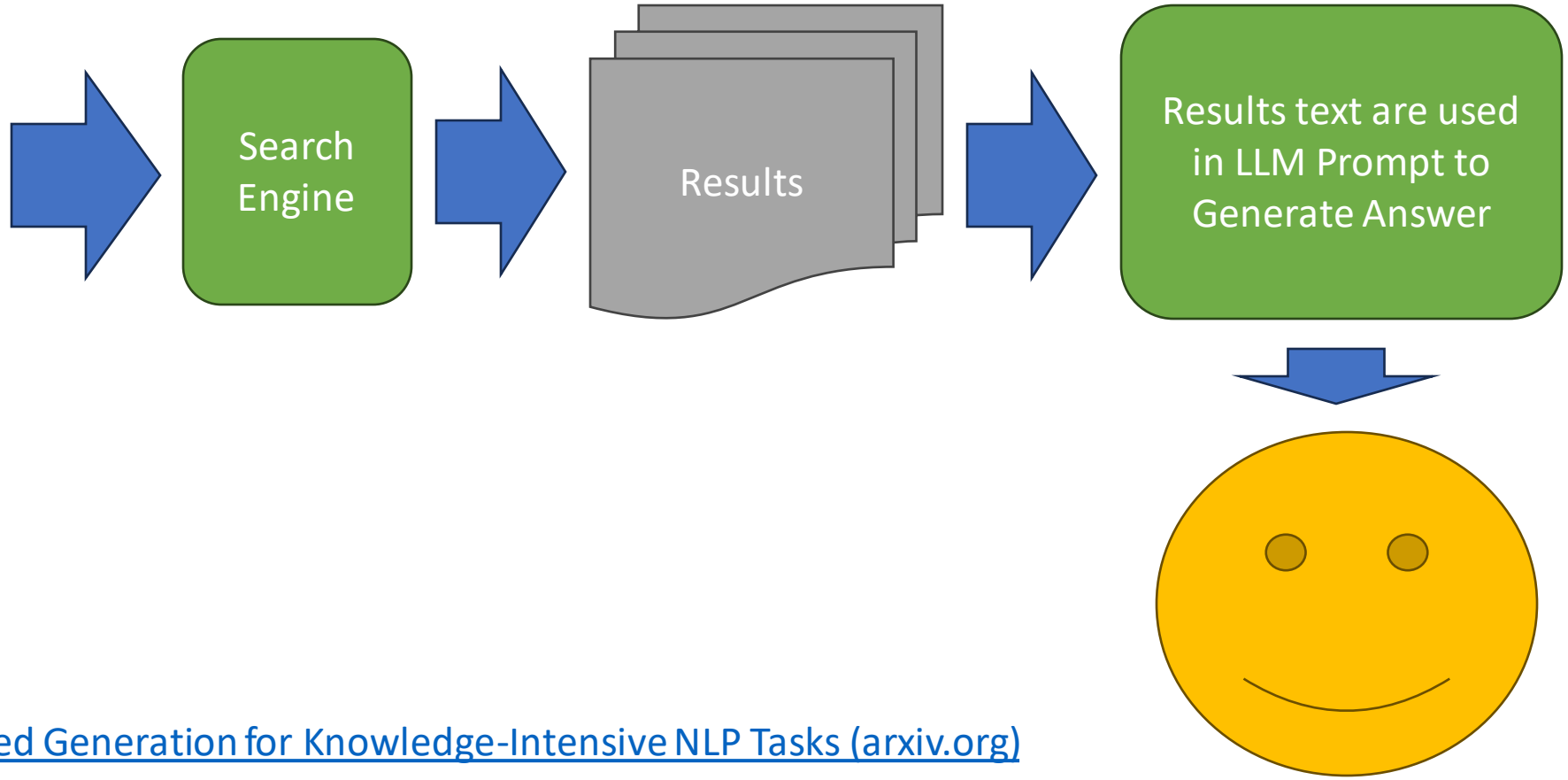
Retrieval Augmented Generation

Question Answering

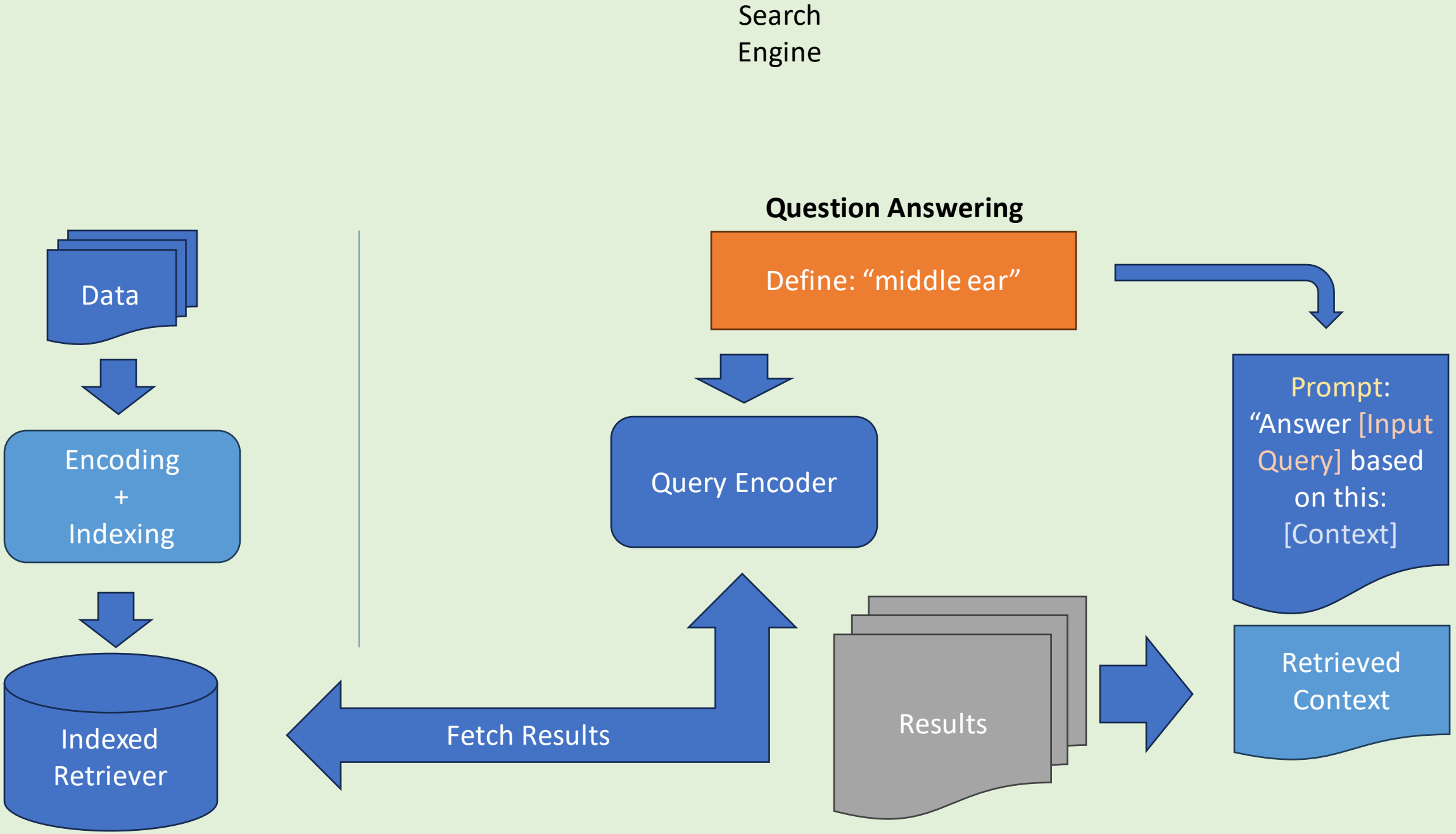
Define: “middle ear”

Fact Verification

Barack Obama was born in Hawaii.



[2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (arxiv.org)


Search
Engine
Question Answering
Data
Define: “middle ear”
Encoding
+
Indexing
Query Encoder
Prompt:
“Answer [Input Query] based on this:
[Context]
Indexed
Retriever
Fetch Results
Results
Retrieved
Context

Retrieval Augmented Generation

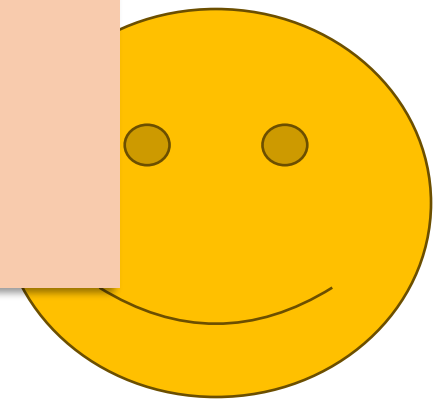
Question Ans

Define: “midd

Fact Verific

Barack Obama w
Hawaii

ts text are used
LM Prompt to
erate Answer

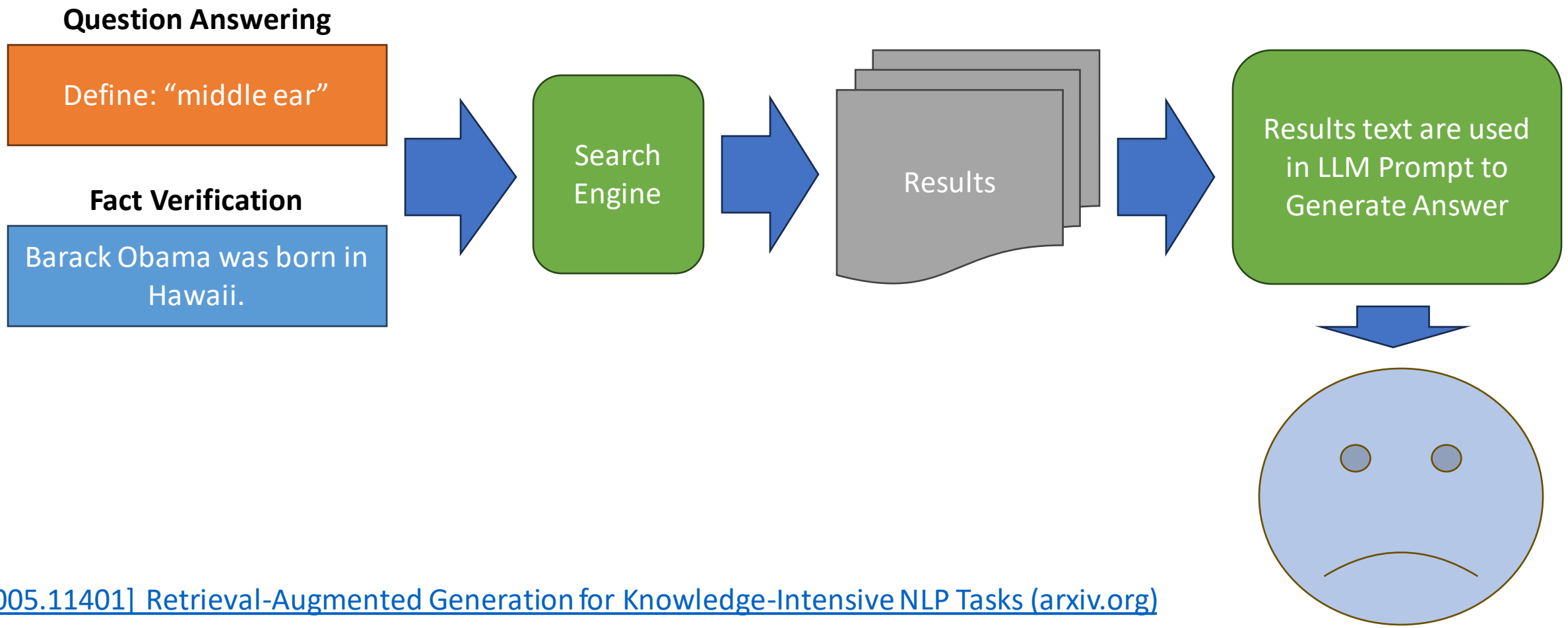


Many Existing Solutions:

- Pinecone’s Canopy
- Llama-Index
- DSPy
- Vespa.ai
- FastRAG
- Haystack
- …

[2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (arxiv.org)

Retrieval Augmented Generation

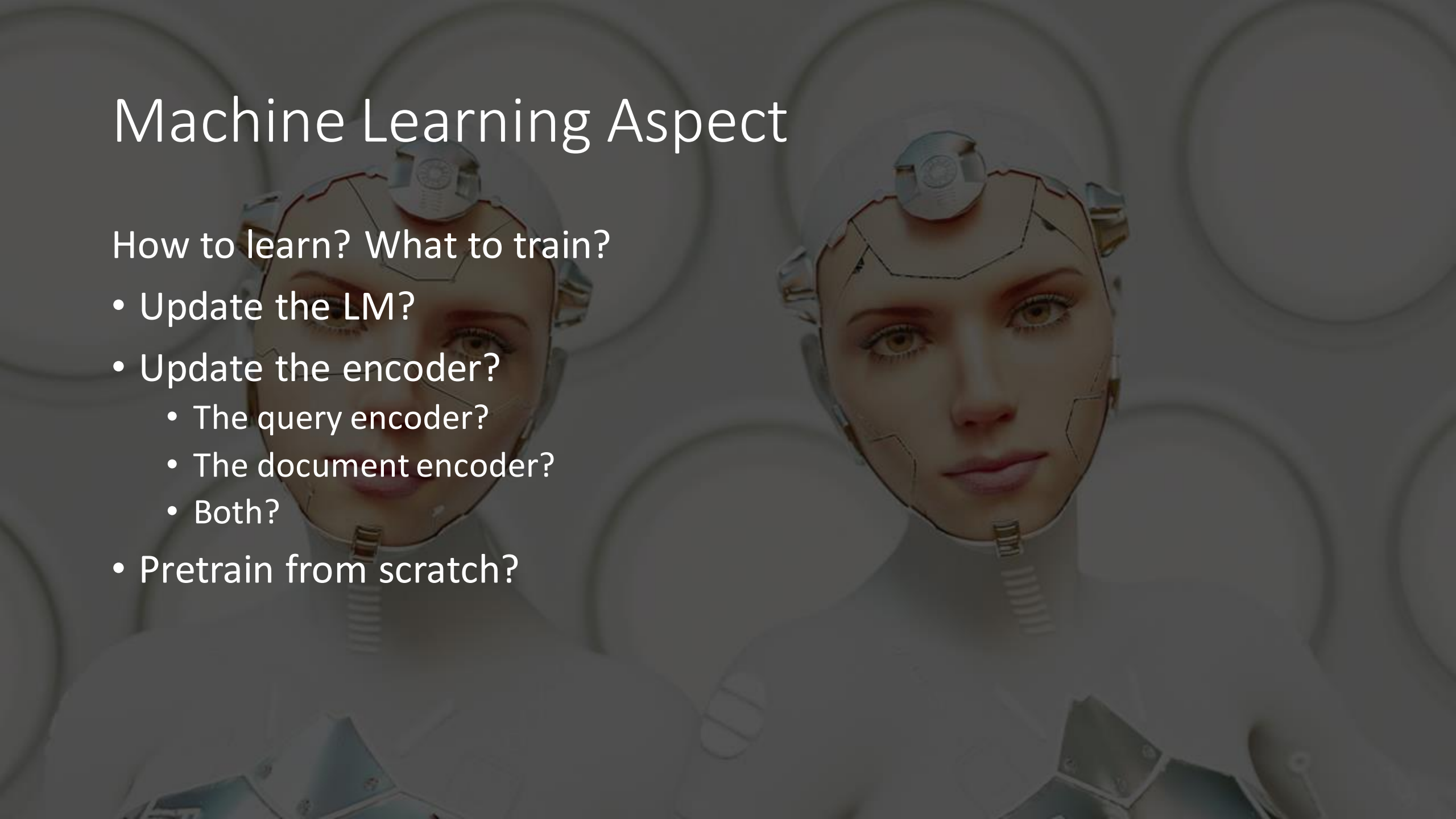


[2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (arxiv.org)

Many Components - Many Questions

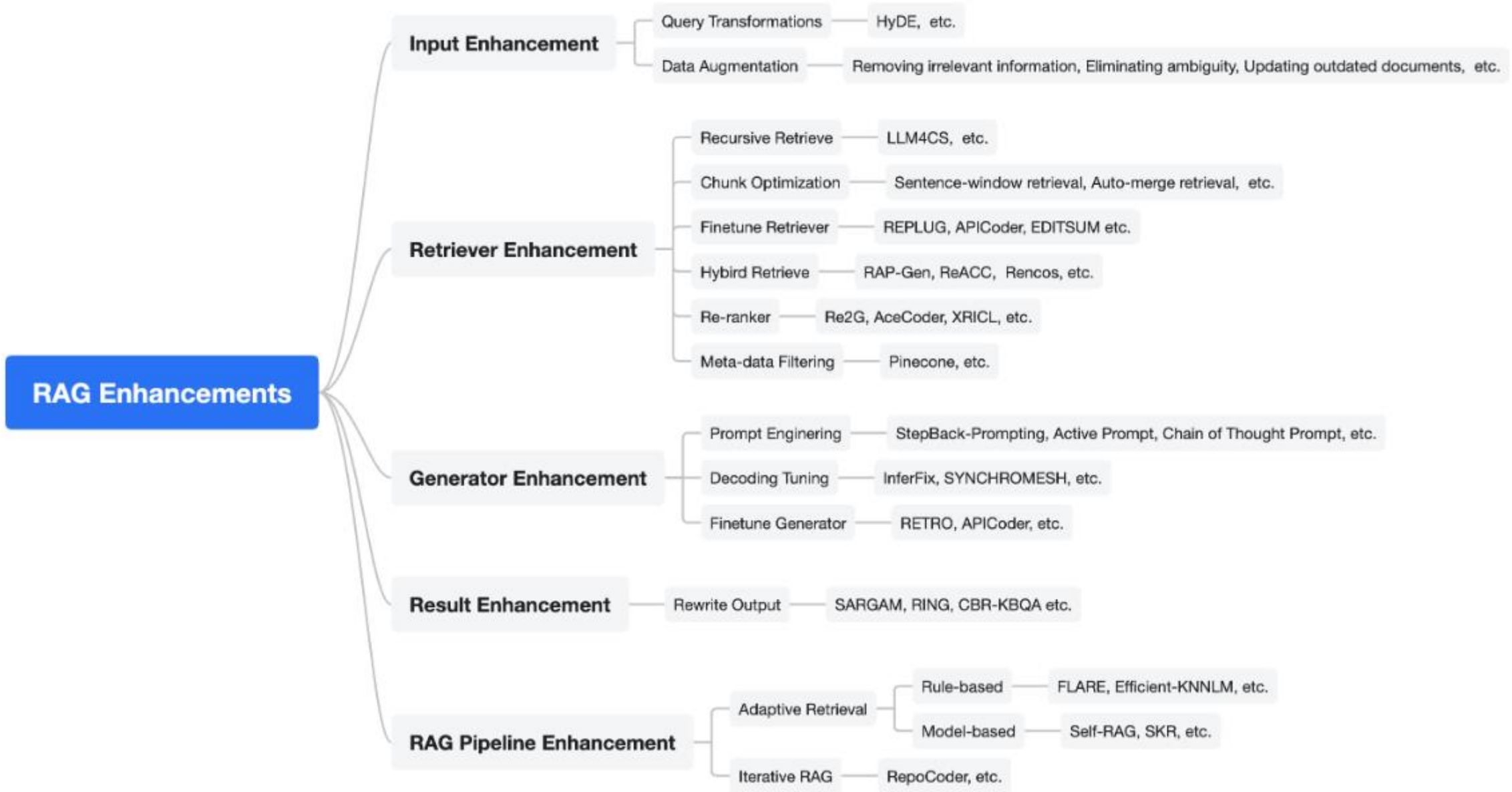
How to index?	How to encode the documents?	How much text is a document?	How to encode the input queries?
How to retrieve?	When to retrieve?	Which prompt to use?	How to pass the context?
What hyperparameters to optimize?	How to post-process the output?	How to verify the output?	How to scale?

Machine Learning Aspect



How to learn? What to train?

- Update the LM?
- Update the encoder?
  - The query encoder?
  - The document encoder?
  - Both?
- Pretrain from scratch?

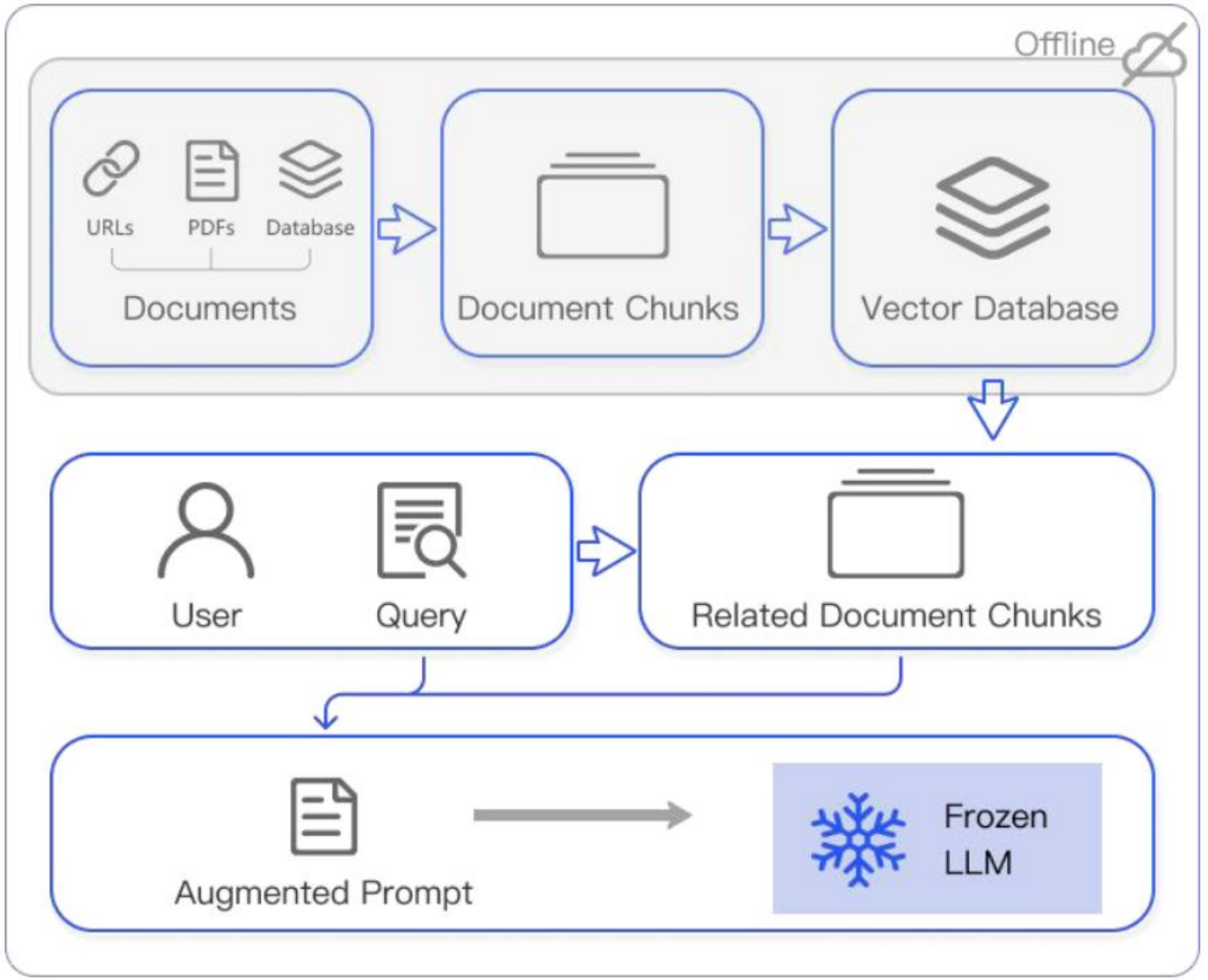

RAG Enhancements
Input Enhancement
Query Transformations
HyDE, etc.
Data Augmentation
Removing irrelevant information, Eliminating ambiguity, Updating outdated documents, etc.
Retriever Enhancement
Recursive Retrieve
LLM4CS, etc.
Chunk Optimization
Sentence-window retrieval, Auto-merge retrieval, etc.
Finetune Retriever
REPLUG, APICoder, EDITSUM etc.
Hybird Retrieve
RAP-Gen, ReACC, Rencos, etc.
Re-ranker
Re2G, AceCoder, XRICL, etc.
Meta-data Filtering
Pinecone, etc.
Generator Enhancement
Prompt Enginering
StepBack-Prompting, Active Prompt, Chain of Thought Prompt, etc.
Decoding Tuning
InferFix, SYNCHROMESH, etc.
Finetune Generator
RETRO, APICoder, etc.
Result Enhancement
Rewrite Output
SARGAM, RING, CBR-KBQA etc.
RAG Pipeline Enhancement
Adaptive Retrieval
Rule-based
FLARE, Efficient-KNNLM, etc.
Model-based
Self-RAG, SKR, etc.
Iterative RAG
RepoCoder, etc.

Naïve RAG/Frozen LLM

1. Indexing
2. Retrieval
3. Generation

No training:

Using pretrained Models.



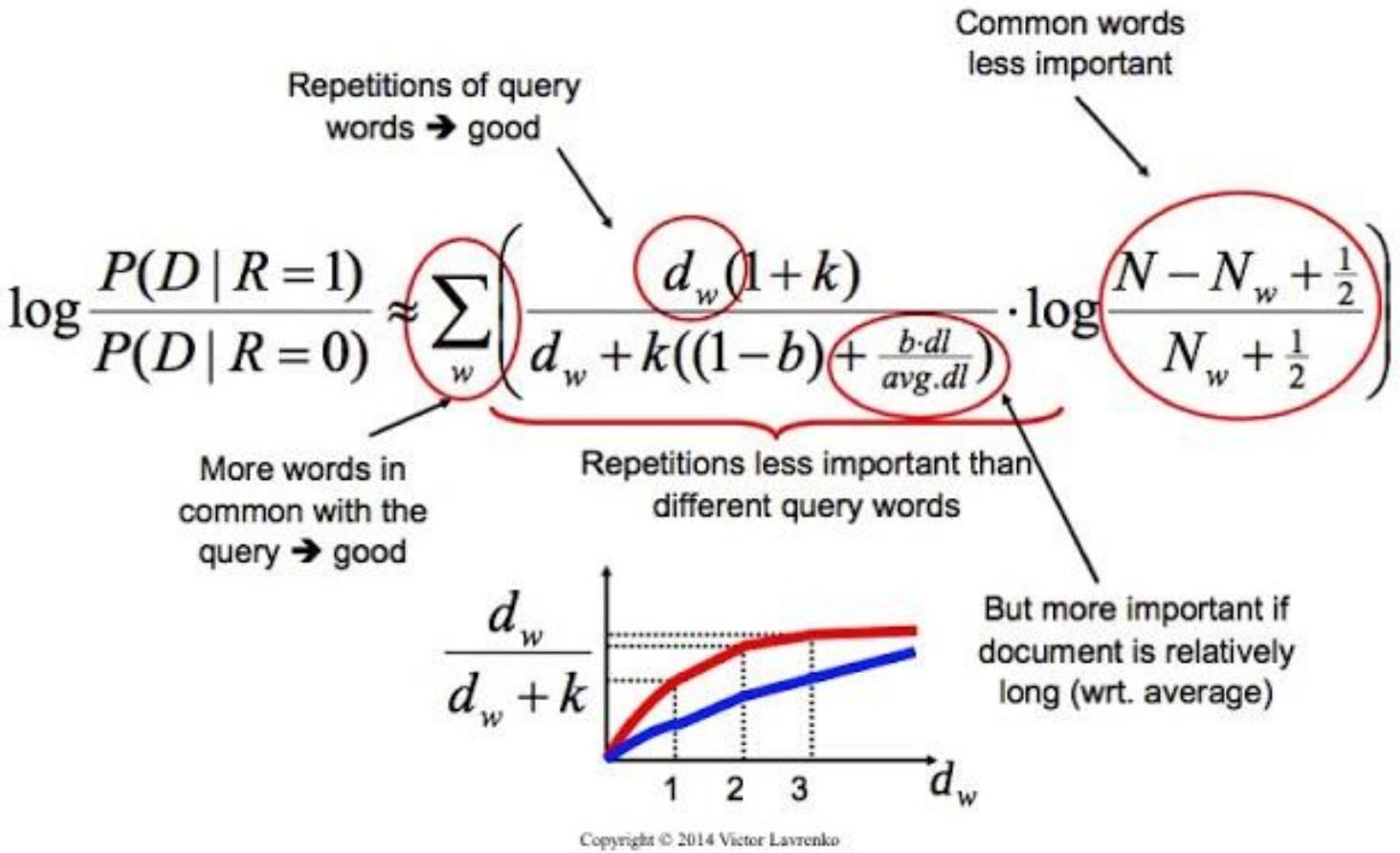
Optimizing Retrieval

- Sparse Retriever
- Dense Retriever
- Other

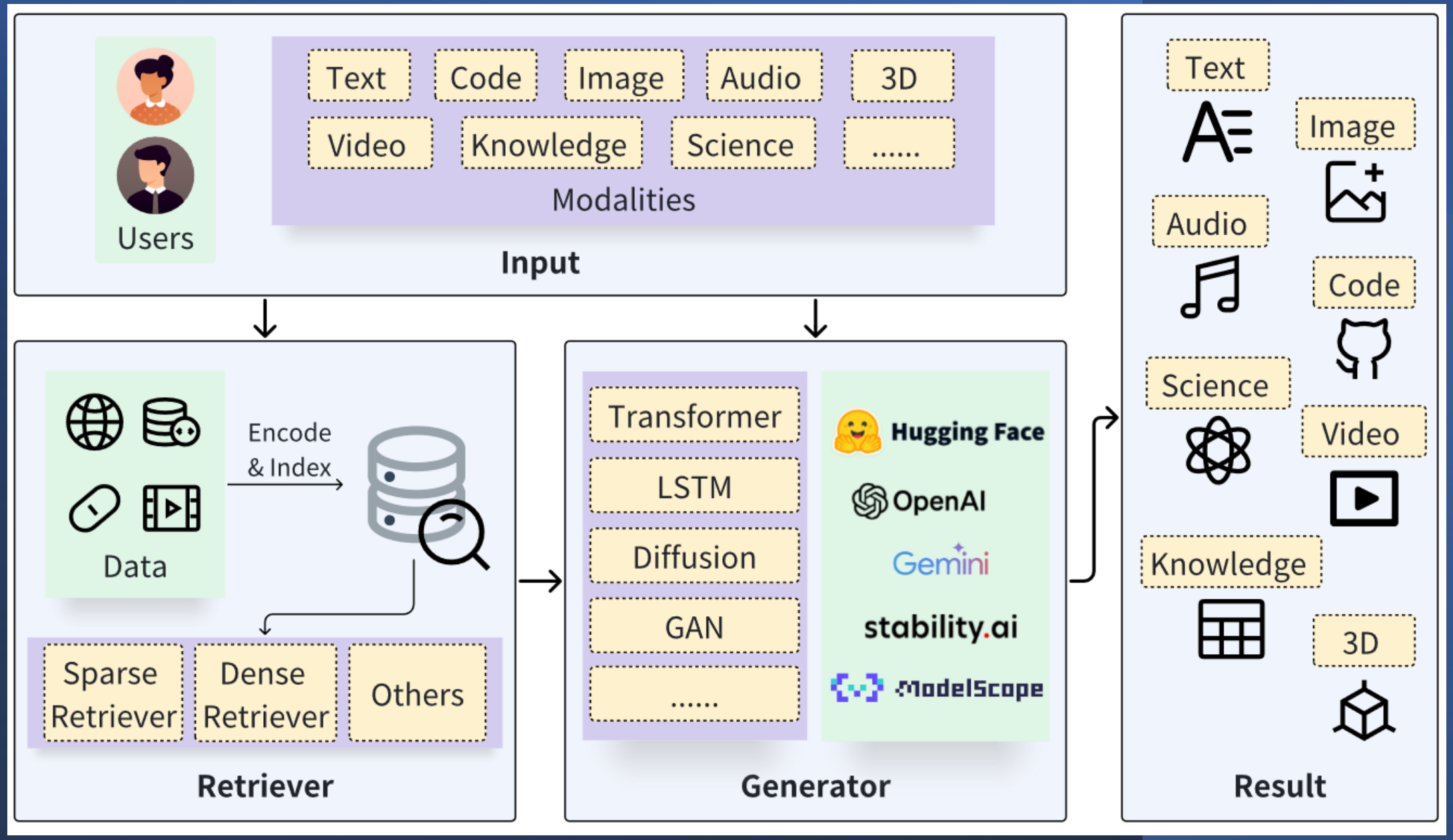
Sparse Retrieval

[0 0 0 0 0 .3 0 0 0 0 0 0 .1 0 ...]

- TF/IDF
- BM25
  - Term Freq. Saturation
  - Avg. Document Length
- LM-Based



- [The Probabilistic Relevance Framework: BM25 and Beyond: Foundations and Trends in Information Retrieval]
- [Document language models, query models, and risk minimization for information retrieval]


Users
Text
Code
Image
Audio
3D
Video
Knowledge
Science
.....
Modalities
Input
Data
Encode
& Index
Sparse
Retriever
Dense
Retriever
Others
Retriever
Transformer
LSTM
Diffusion
GAN
.....
Hugging Face
OpenAI
Gemini
stability.ai
ModelScope
Generator
Text
Image
Audio
Code
Science
Video
Knowledge
3D
Result

Dense Retriever

Documents embedded as a dense vector

BERT / RoBERTa

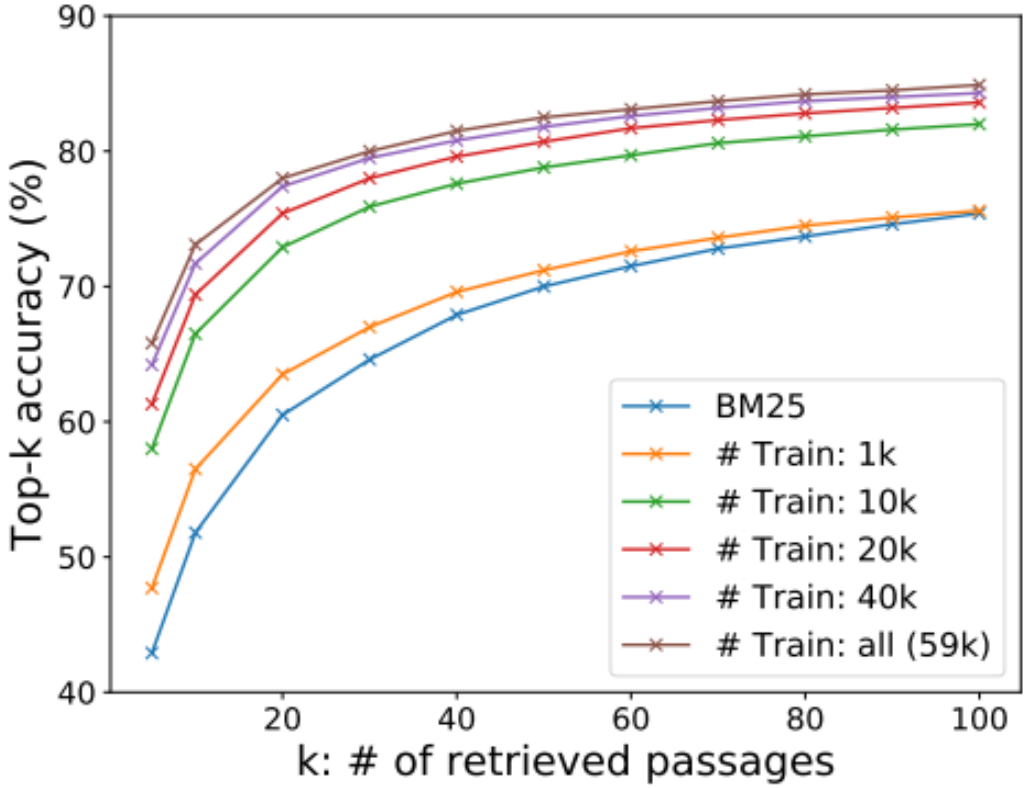
Approximate Nearest Neighbor (ANN) index

Cosine similarity

Inner product

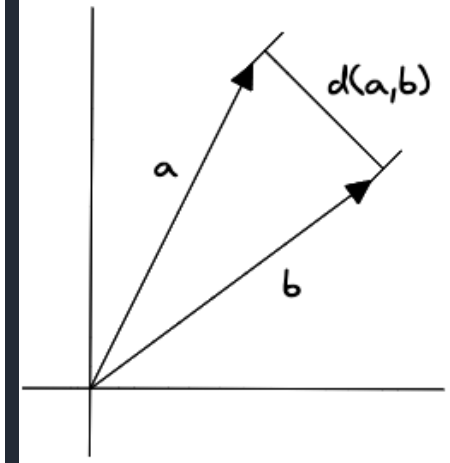
L2-Distance

Captures Semantic Similarity, synonyms

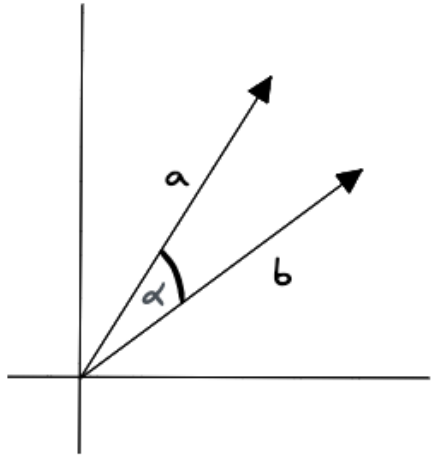


[2004.04906] Dense Passage Retrieval for Open-Domain Question Answering (arxiv.org) (OrQA)

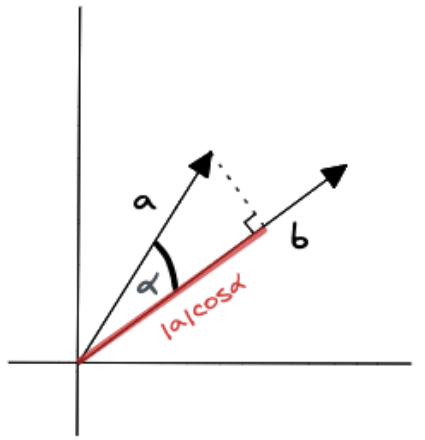
Dense Retriever – Vector Databases



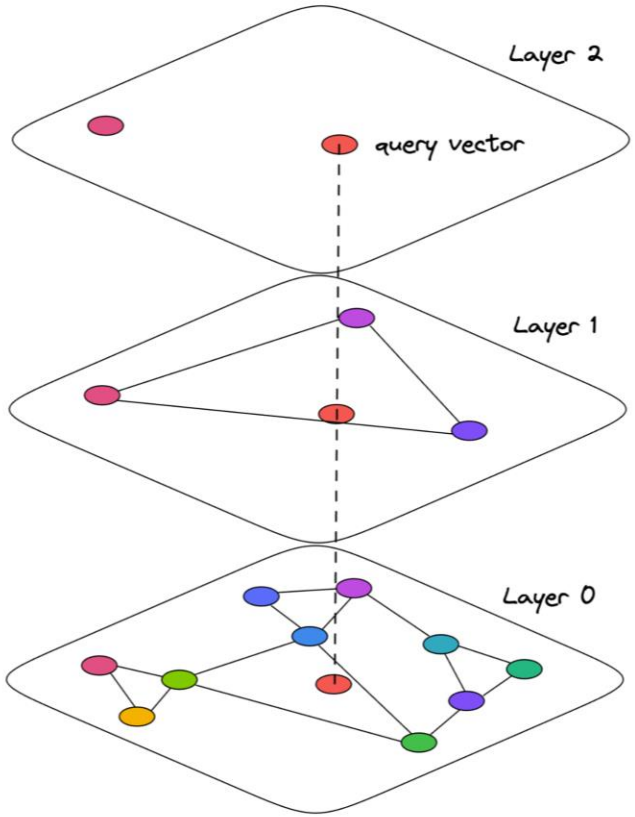
Euclidean Distance



Cosine Similarity



Dot Product

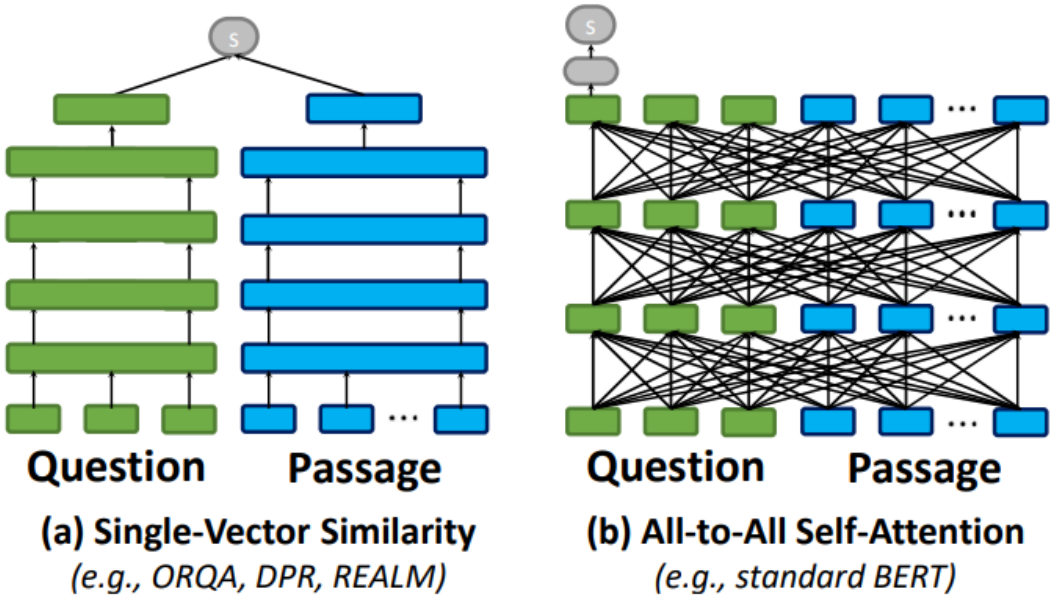


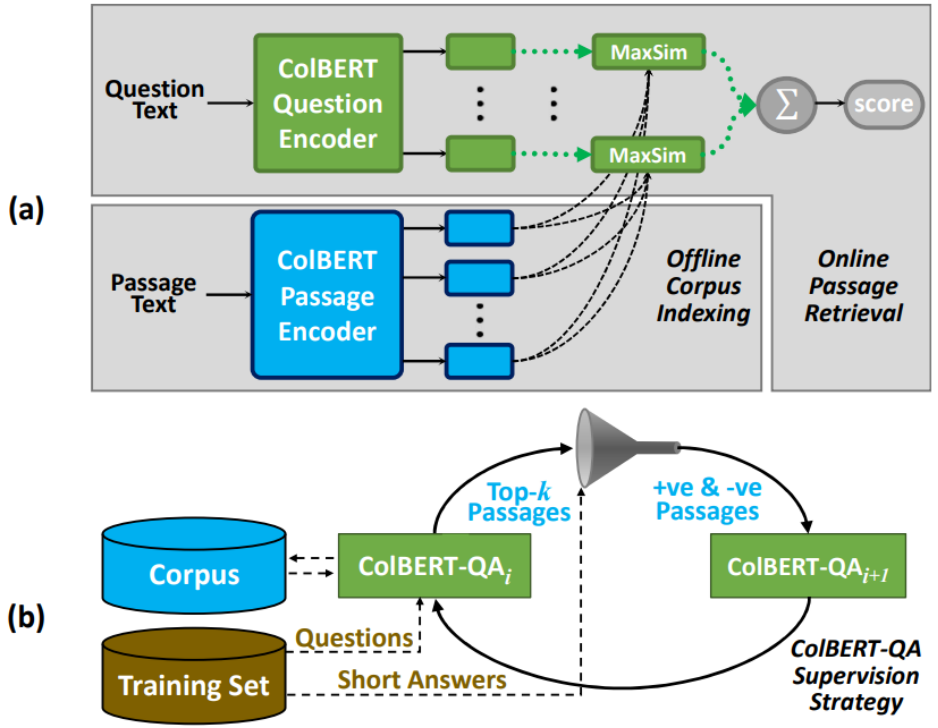
- Fast *top-k* vector similarity search - on GPU

- Example Databases (ANN Search Libraries):
  - FAISS
  - Vector.dev
  - q-drant
  - Chroma
  - Milvus

Dense Retriever – Beyond Dot-Product

Similarity using Siamese networks





[2007.00814] Relevance-guided Supervision for OpenQA with ColBERT (arxiv.org)

ColBERT

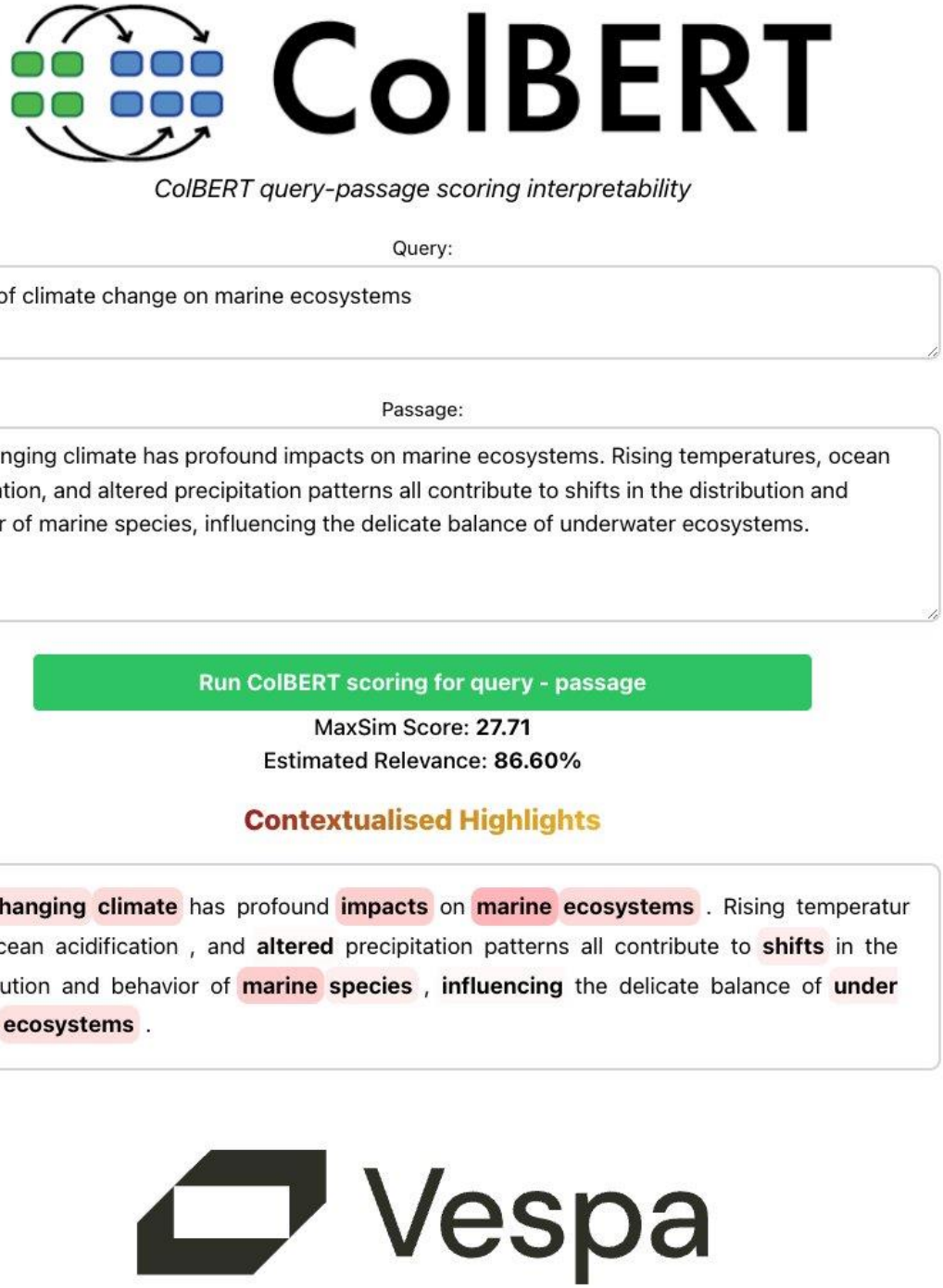
Not a single document-vector:

a set of token-vectors

Enables better matching between Query and Document

- ColBERT v2:
  - Distilled mode
  - vector compression

[ColBERT Inference in the Browser (aiserv.cloud)]



Vespa

Retrieval Types – A Comparison

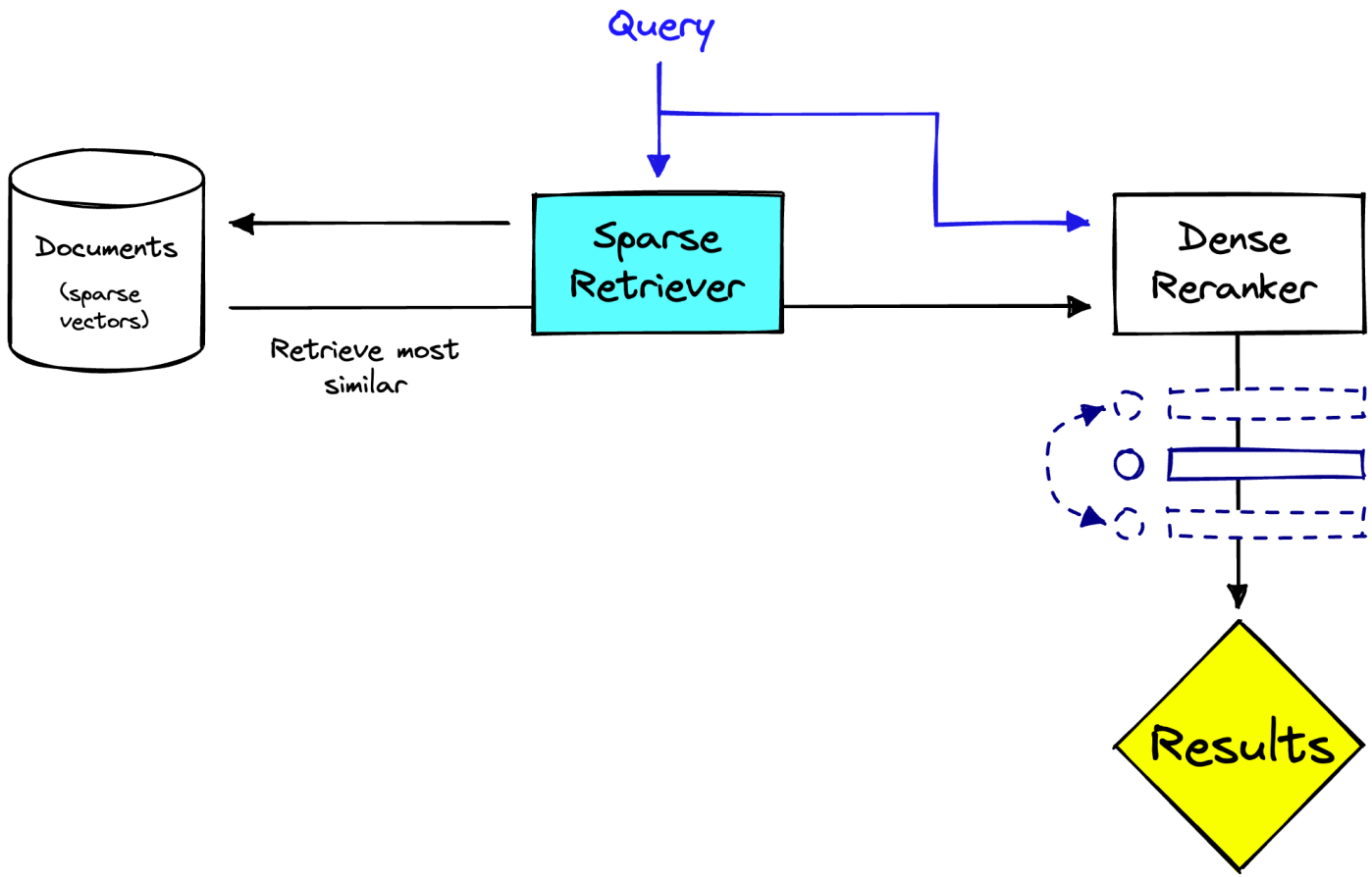
Sparse

- Pros:
  - Allows fast retrieval
  - No fine-tuning required
  - Exact term matching
  - Works well for long documents
  - Good baseline
- Cons:
  - Vocabulary Mismatch Problem
  - No Semantic similarity
  - Performance quite fixed

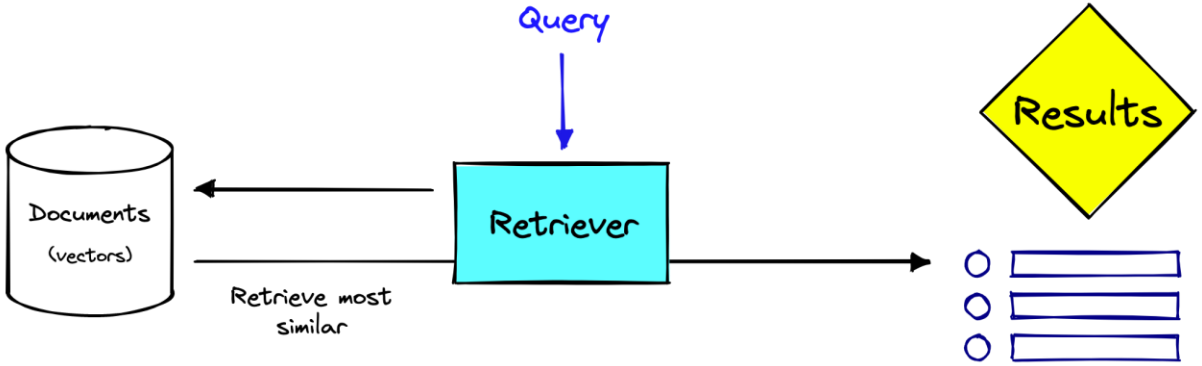
Dense

- Pros:
  - Multi-modal (and cross-modal)
  - Allow semantic similarity
  - Performs well (with training)
  - Multi-lingual
- Cons:
  - Requires fine-tuning (with a labeled training-set)
  - Can’t generalize between domains
  - Require more compute & memory
  - Don’t allow exact match
  - interpretability

Sparse Retriever + Dense Re-ranker

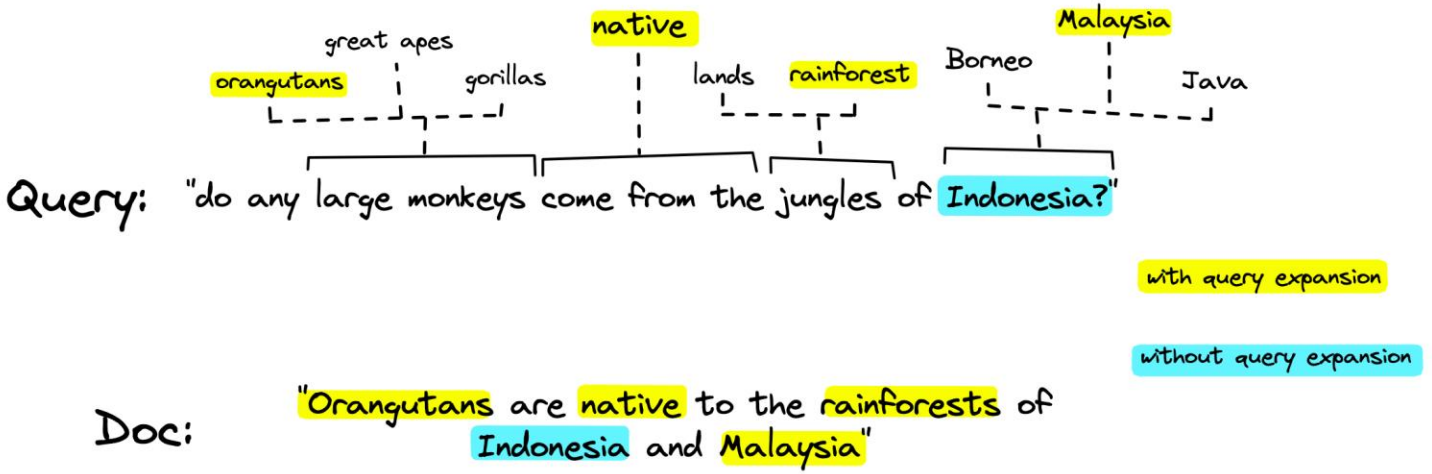


Retrieval: SotA



- SPLADE — the **Sp**arse **L**exical **a**n**d** **E**xpansion model
  - Query Expansion using LM distribution
  - A semi-sparse vector with synonyms & word variations

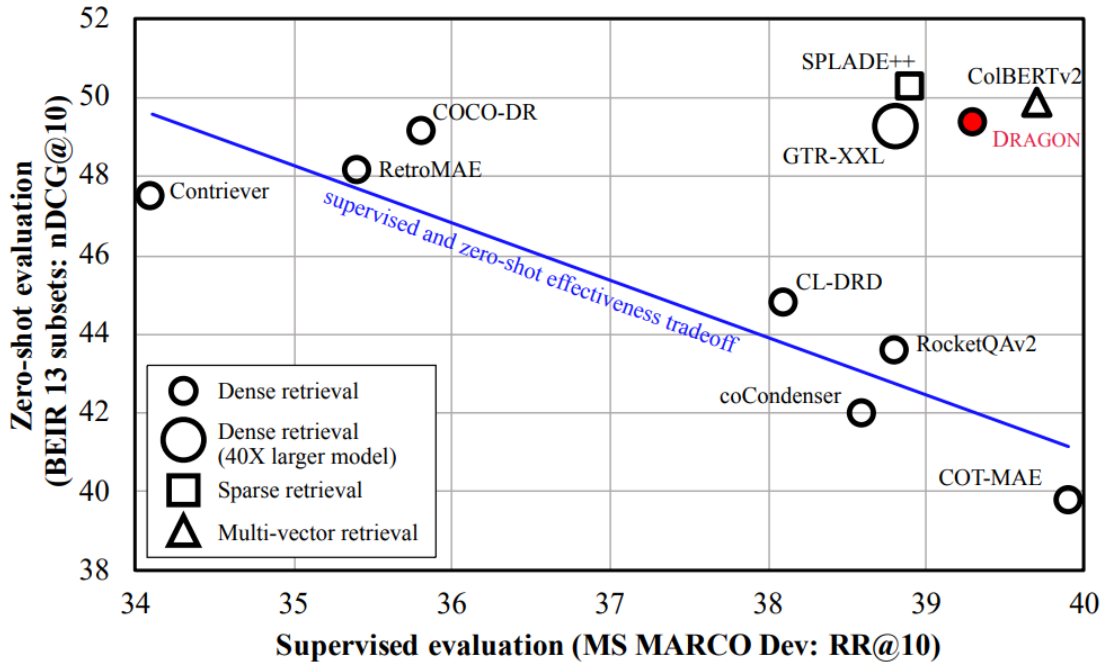
- SPLADE V2
  - Max-pooling
  - Model Distillation
- SPLADE++
  - Negative Sampling



- SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking
- SPLADE v2: Sparse Lexical and Expansion Model for Information Retrieval

Retrieval: SotA

- Zero-Shot Transfer Learning works well between **tasks**, but less between **domains** (different vocabulary)

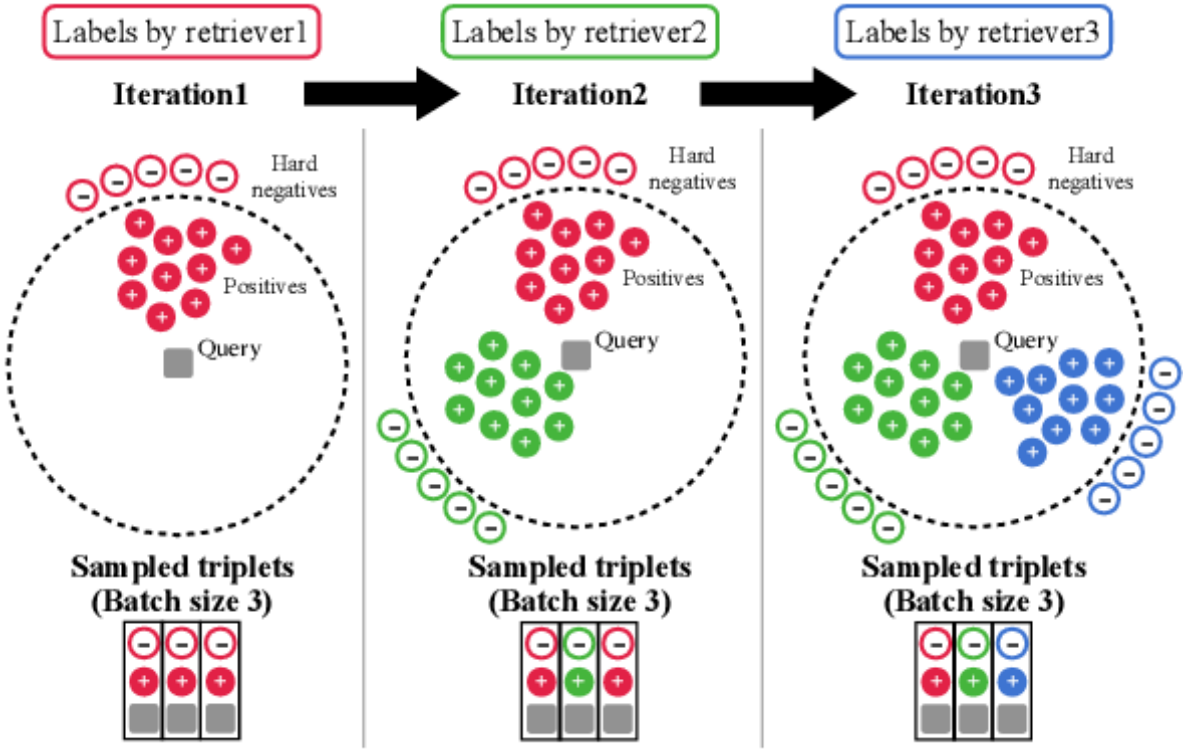


How to Train Your Dragon: Diverse Augmentation Towards Generalizable Dense Retrieval

Retrieval: SotA

- Dragon:
  Dense **R**etriever trained with diverse **A**u**G**mentati**ON**

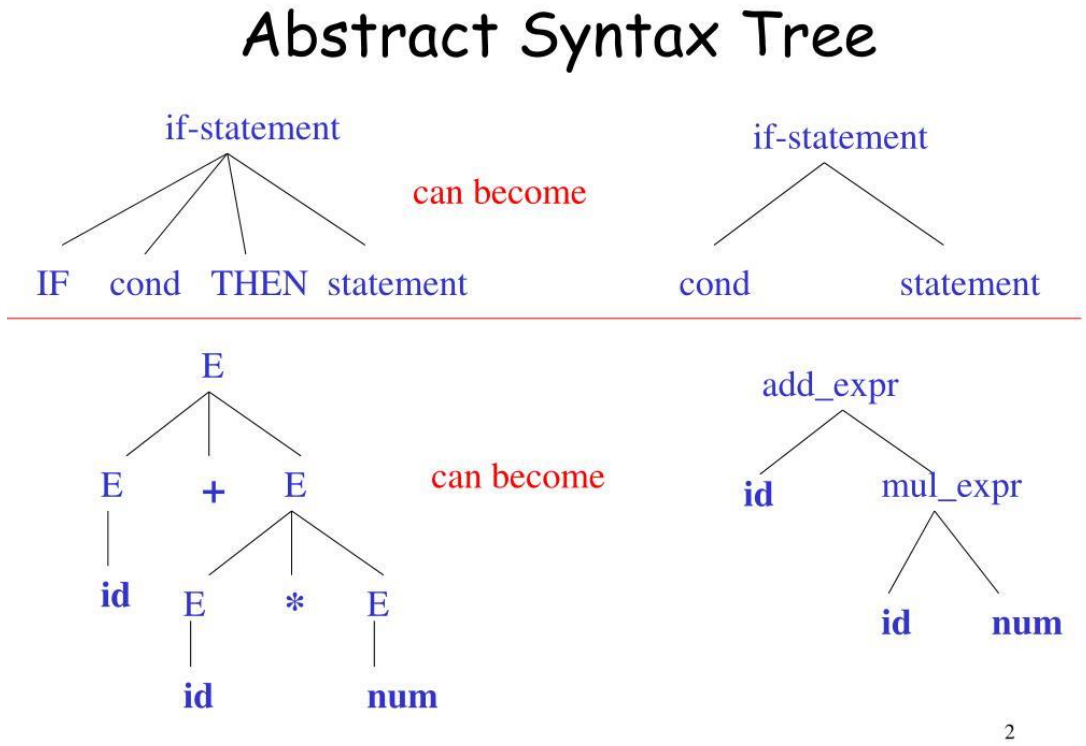
- Hybrid taining a DR with iterative augmentation from:
  - Sparse Retriever
  - Dense Retriever
  - Multi-Vector Retriever
- Positive + Hard-Negative Sampling
- 28m queries; 5 Days on A100
- Available on HF



[How to Train Your Dragon: Diverse Augmentation Towards Generalizable Dense Retrieval]

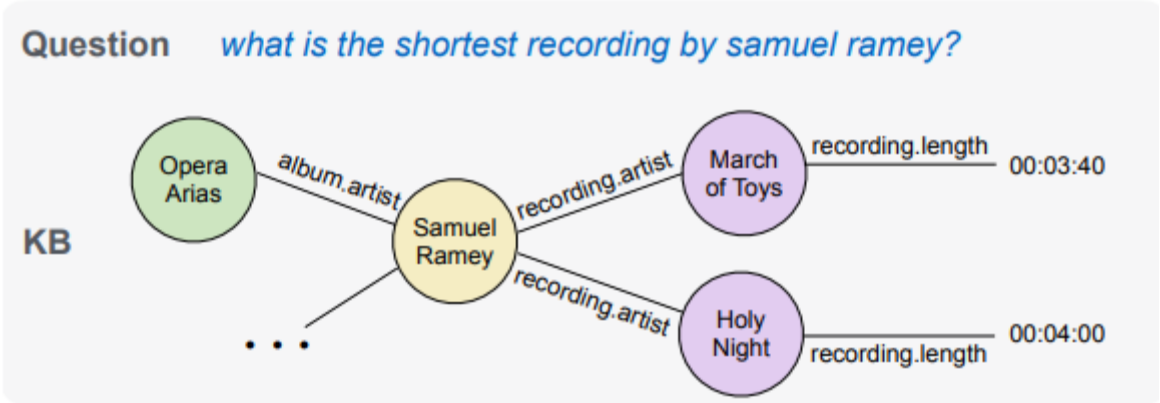
Side-Note: Other Retrievers

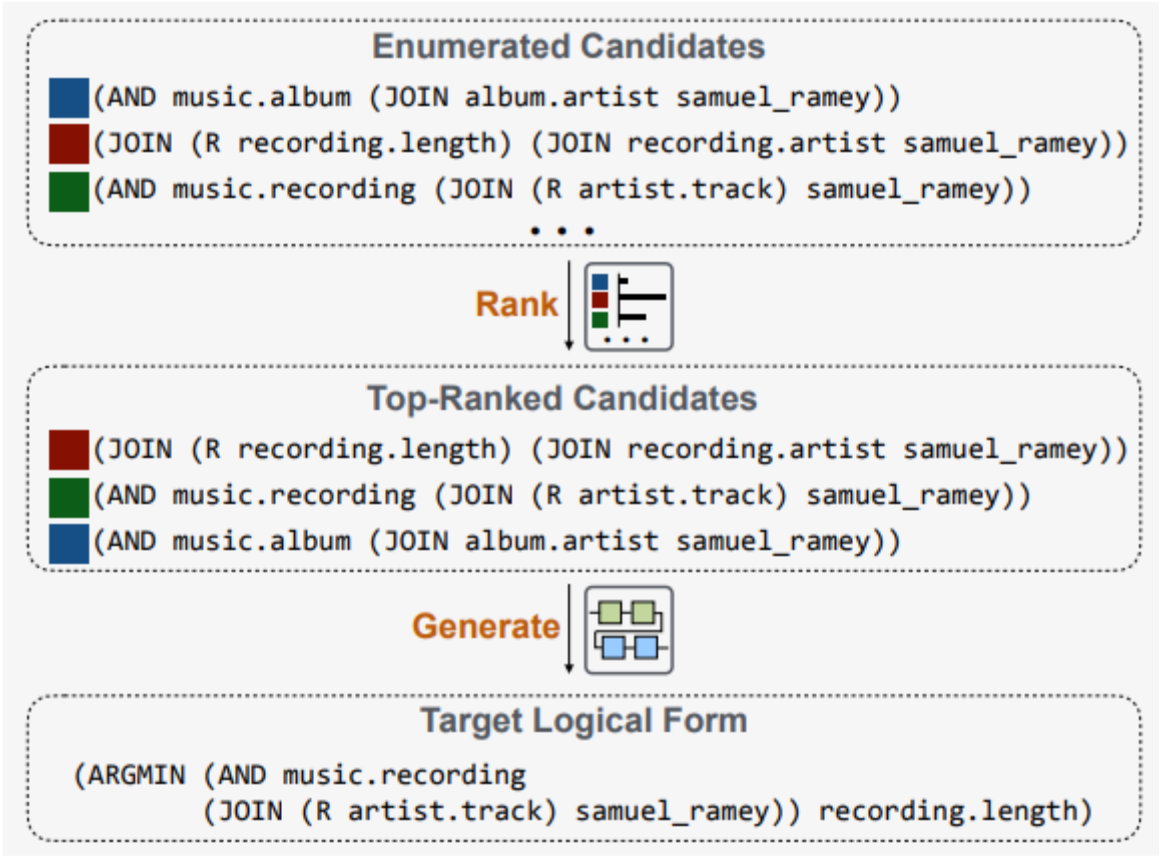
- Code:
  - Abstract Syntax Trees
    ReCode: Retrieval-Based Neural Code Generation (arxiv.org)
  - Target Similarity Tuning (**TST**): Synchromesh: Reliable code generation from pre-trained language models (arxiv.org)



Other Retrievers

- Knowledge Base:
  - Rank And Generate: [RnG-KBQA: Generation Augmented Iterative Ranking for Knowledge Base Question Answering (arxiv.org)]
  - ECBRF: [End-to-end Case-Based Reasoning for Commonsense Knowledge Base Completion - ACL Anthology]



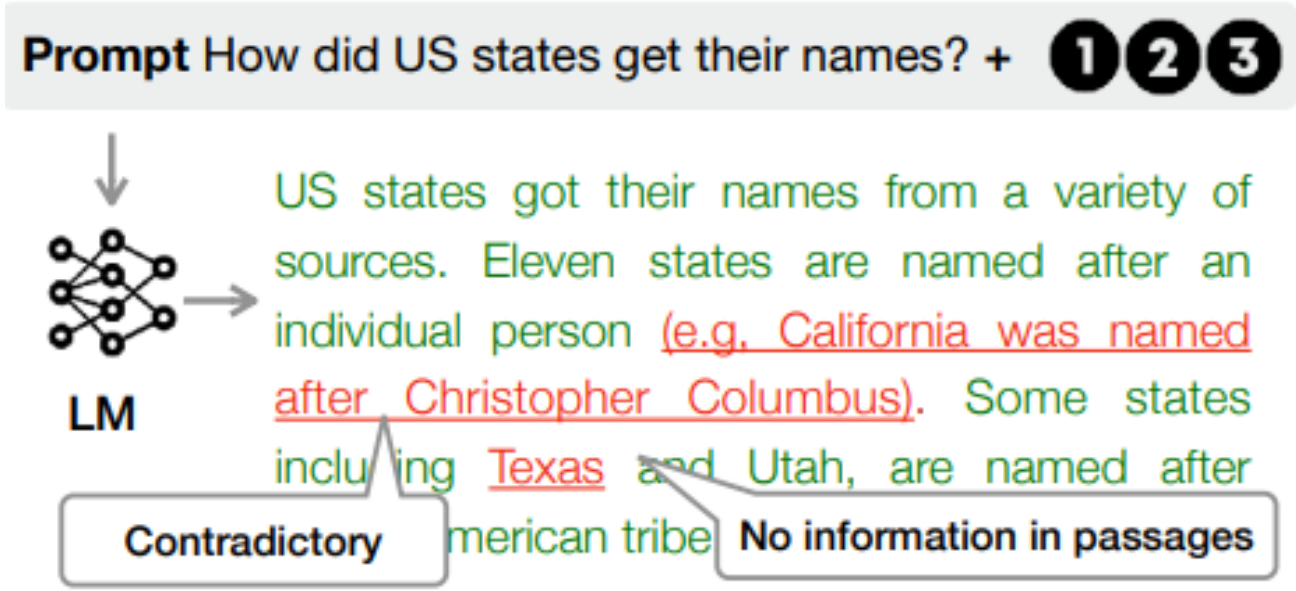


Self-Control

When the retrieved documents are wrong

Critical Models – Self-Rag

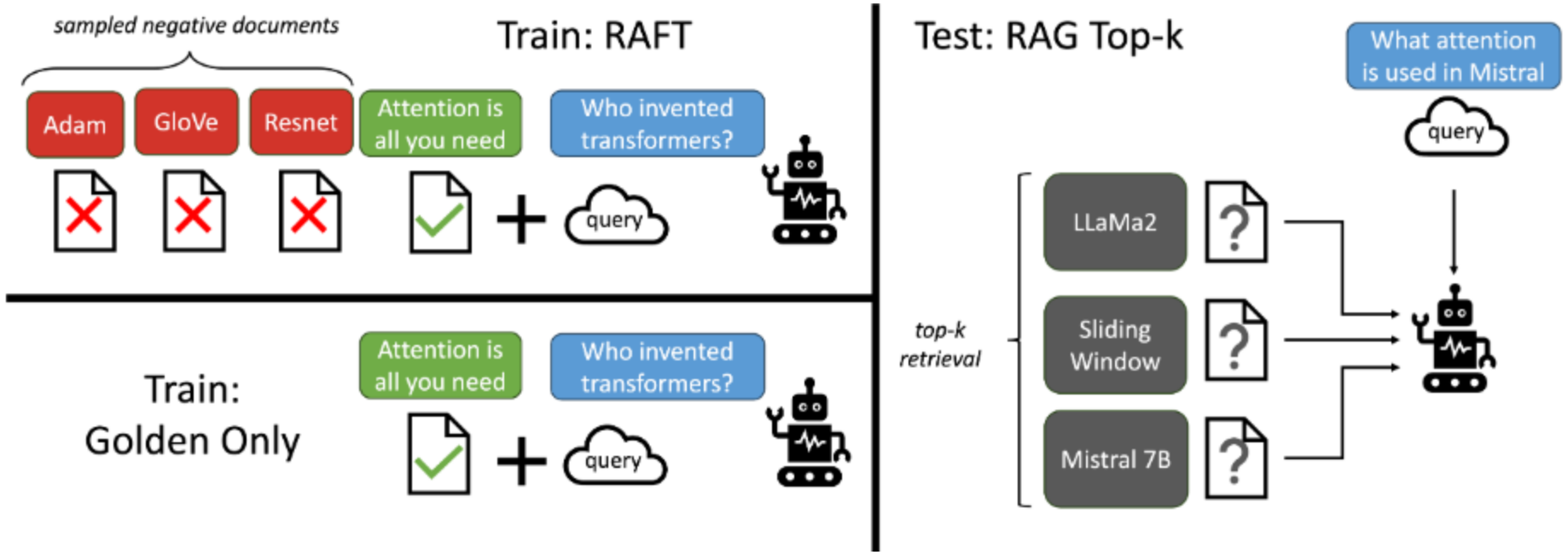
Retrains the LM to evaluate the retrieved content to be relevant to the query:



Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection (arxiv.org)

Critical Models - RAFT

Refines LLM with distracting negative documents and Chain of Thoughts (CoT)



[2403.10131] RAFT: Adapting Language Model to Domain Specific RAG (arxiv.org)

Critical Models - RAFT

Question: The Oberoi family is part of a hotel company that has a head office in what city?

context: [The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group]...[It is located in city center of Jakarta, near Mega Kuningan, adjacent to the sister JW Marriott Hotel. It is operated by The Ritz-Carlton Hotel Company. The complex has two towers that comprises a hotel and the Airlangga Apartment respectively]...[The Oberoi Group is a hotel company with its head office in Delhi.]

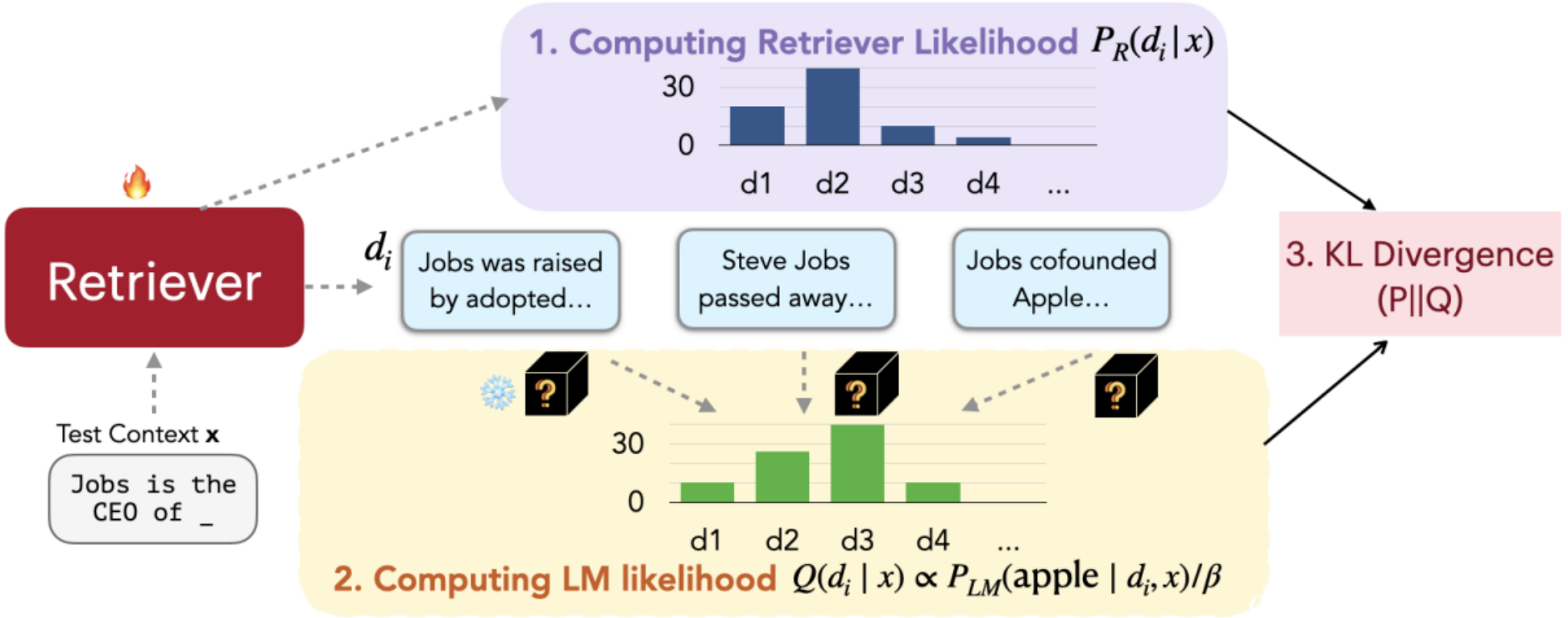
Instruction: Given the question, context and answer above, provide a logical reasoning for that answer. Please use the format of: ##Reason: {reason} ##Answer: {answer}.

CoT Answer: ##Reason: The document ##begin_quote## The Oberoi family is an Indian family that is famous for its involvement in hotels, namely through The Oberoi Group. ##end_quote## establishes that the Oberoi family is involved in the Oberoi group, and the document ##begin_quote## The Oberoi Group is a hotel company with its head office in Delhi. ##end_quote## establishes the head office of The Oberoi Group. Therefore, the Oberoi family is part of a hotel company whose head office is in Delhi. ##Answer: Delhi

[2403.10131] RAFT: Adapting Language Model to Domain Specific RAG (arxiv.org)

Critical Models – RePlug (+LSR)

Uses Black-Boxed LLMs to adapt retrievers



[2301.12652] REPLUG: Retrieval-Augmented Black-Box Language Models (arxiv.org)

Critical Models - RALM

AI21 labs

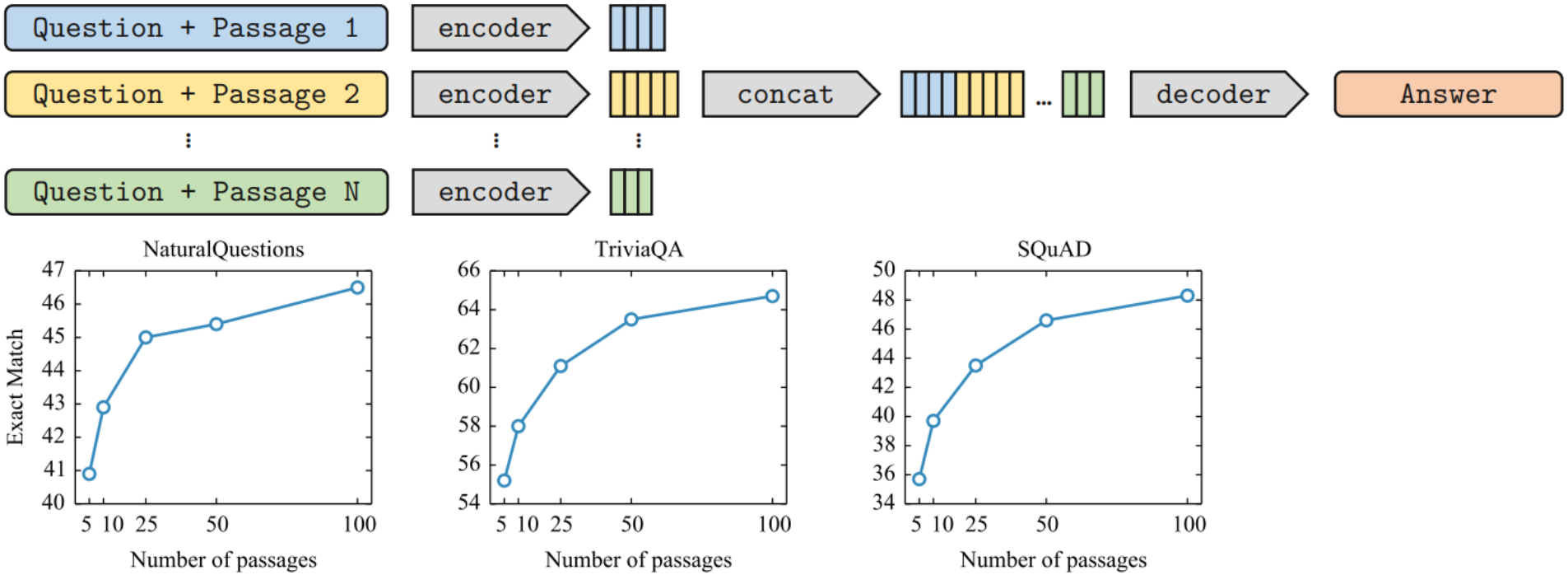
- **R**etrieval-**A**ugmented **L**anguage **M**odeling
- Frozen RAG with BM25
- Training a predictive **re-ranker** to filter the retrieved documents

[In-Context Retrieval-Augmented Language Models - ACL Anthology]

Contextualization of Retriever & Generator

FiD

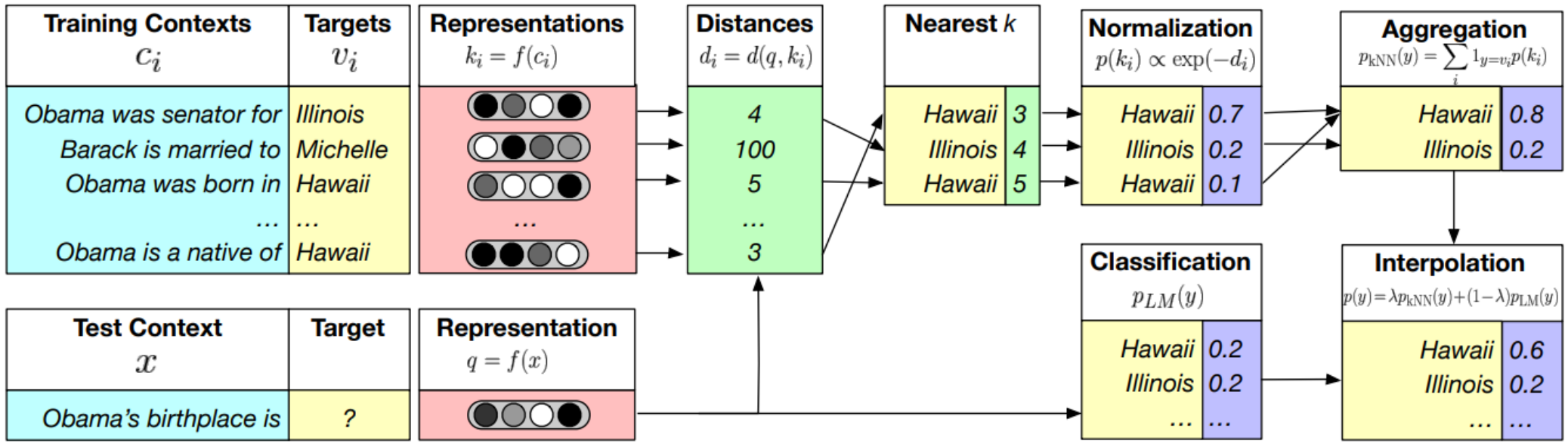
- Overcomes the document limitation of the original RAG
- Encoding the passages, concatenating, decoding into an answer



[2007.01282] Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering (arxiv.org)

*k*NN LM

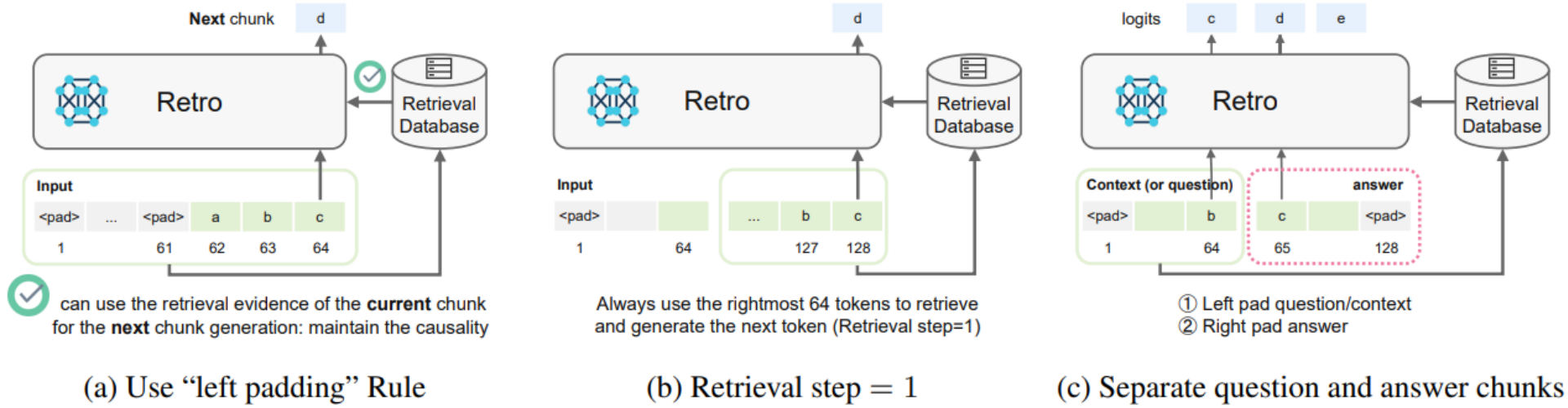
- Cluster the “next-token” with kNN



[1911.00172] Generalization through Memorization: Nearest Neighbor Language Models (arxiv.org)

Retro++

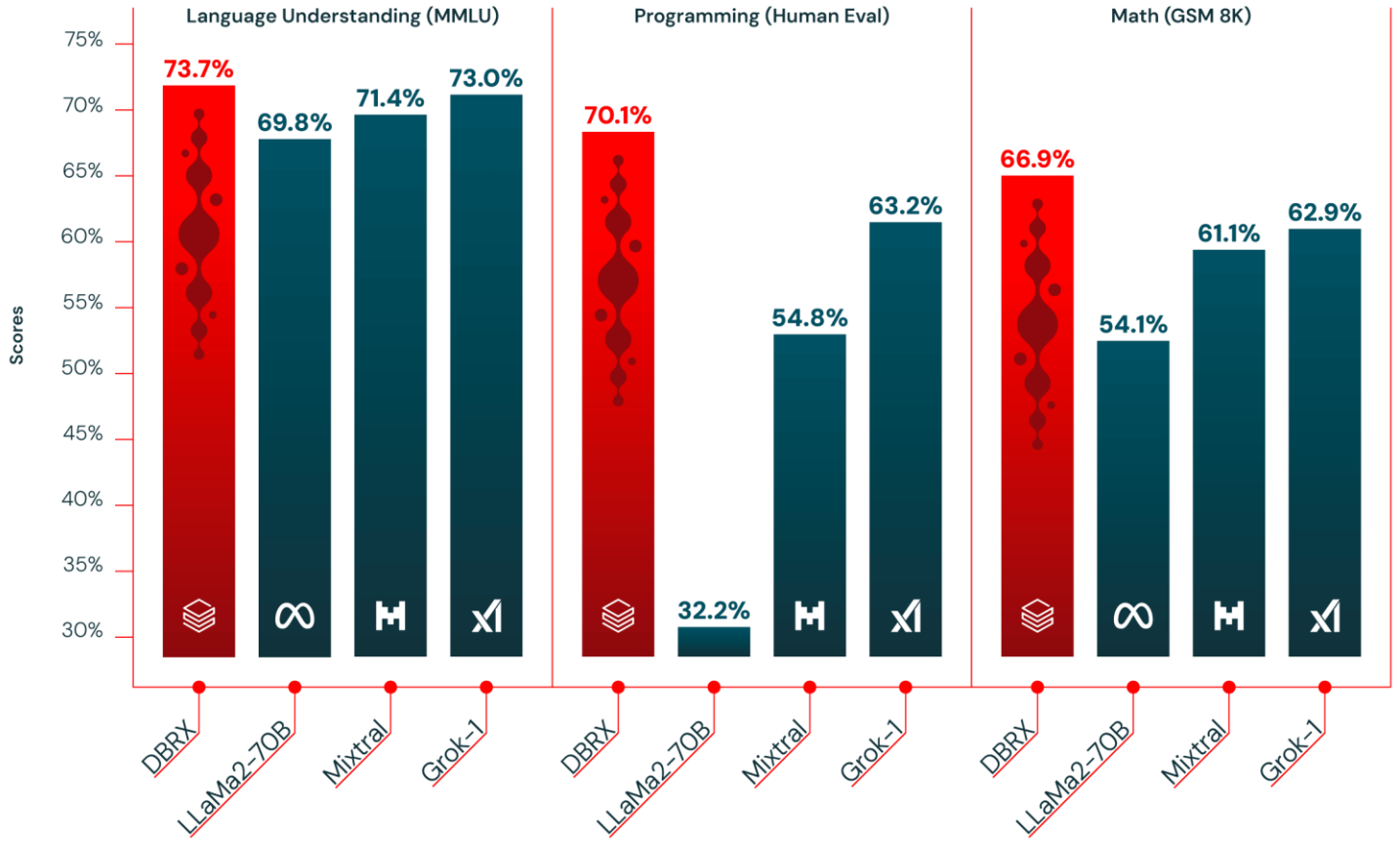
- An open version of the original *Retro* (DeepMind)
- Decoder Only LM
- Performs retrieval ***during*** answer generation



[2304.06762] Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study (arxiv.org)

FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

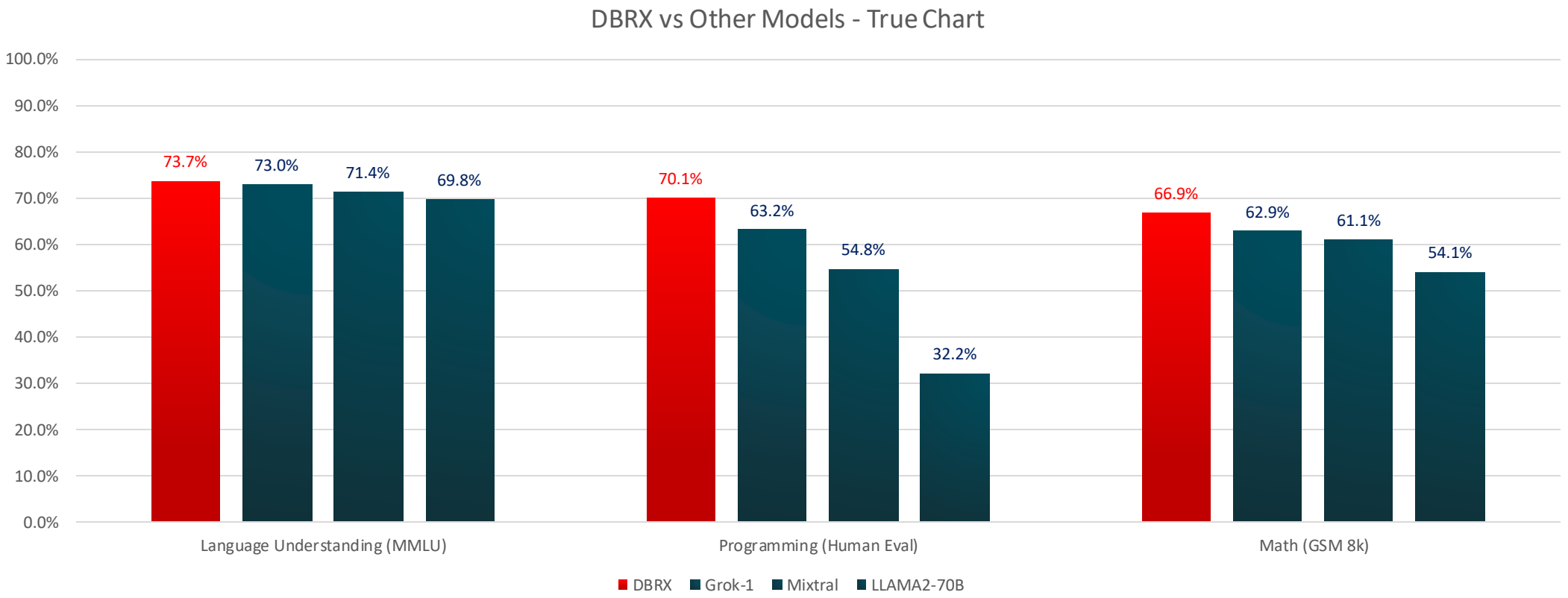
DBRX (formerly MosaicML)



[Introducing DBRX: A New State-of-the-Art Open LLM | Databricks]

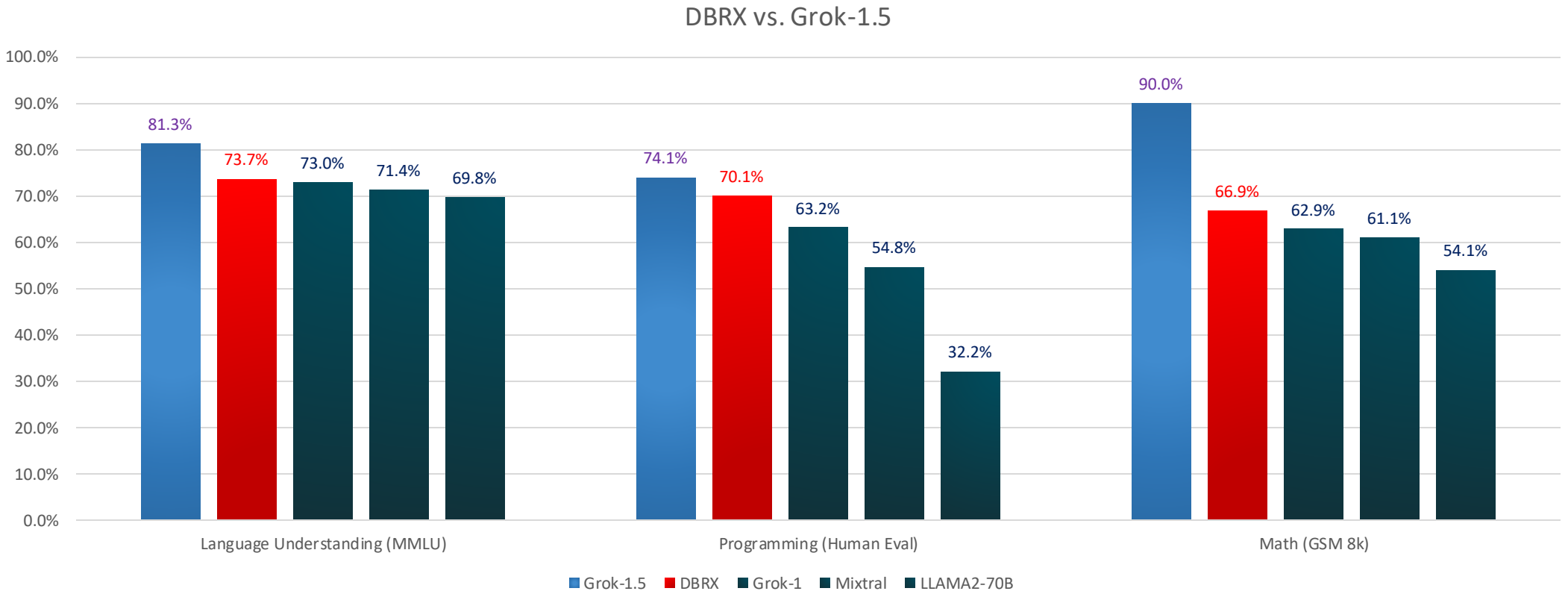
FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

DBRX – Correct Perspective



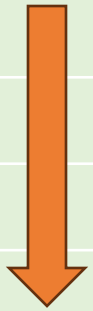
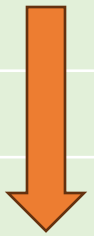
FH CAMPUS WIEN
UNIVERSITY OF APPLIED SCIENCES

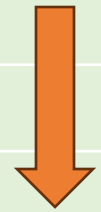
DBRX vs. Grok-1.5



DBRX – Open-Source **132B** Parameters

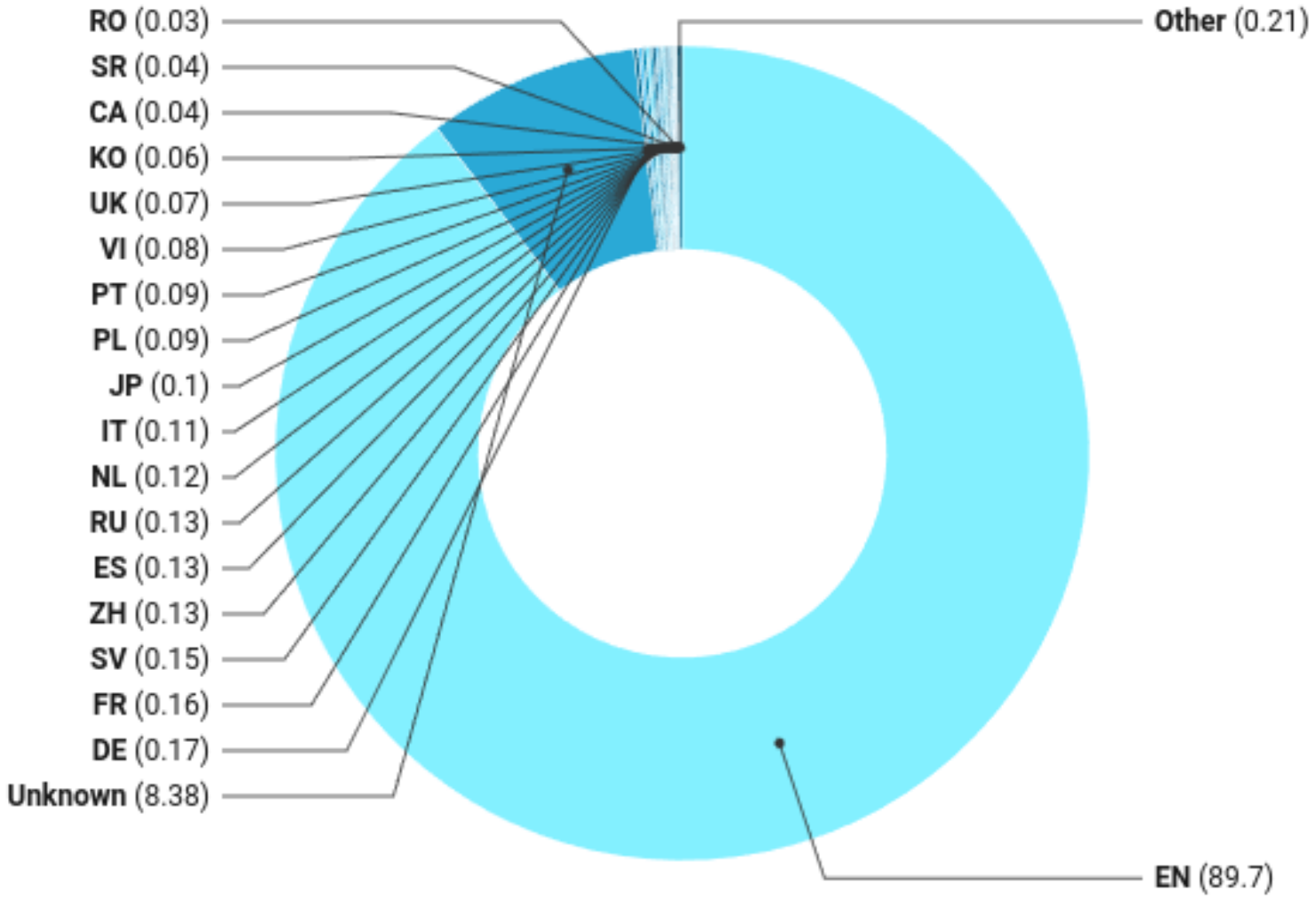
Model	DBRX Instruct	Mixtral Instruct	GPT-3.5 Turbo (API)	GPT-4 Turbo (API)
Answer in Beginning Third of Context	45.1%	41.3%	37.3%*	**49.3%**
Answer in Middle Third of Context	45.3%	42.7%	37.3%*	**49.0%**
Answer in Last Third of Context	48.0%	44.4%	37.0%*	**50.9%**
2K Context	59.1%	64.6%	36.3%	**69.3%**
4K Context	**65.1%**	59.9%	35.9%	63.5%
8K Context	59.5%	55.3%	45.0%	**61.5%**
16K Context	27.0%	20.1%	**31.7%**	26.0%
32K Context	19.9%	14.0%	—	**28.5%**



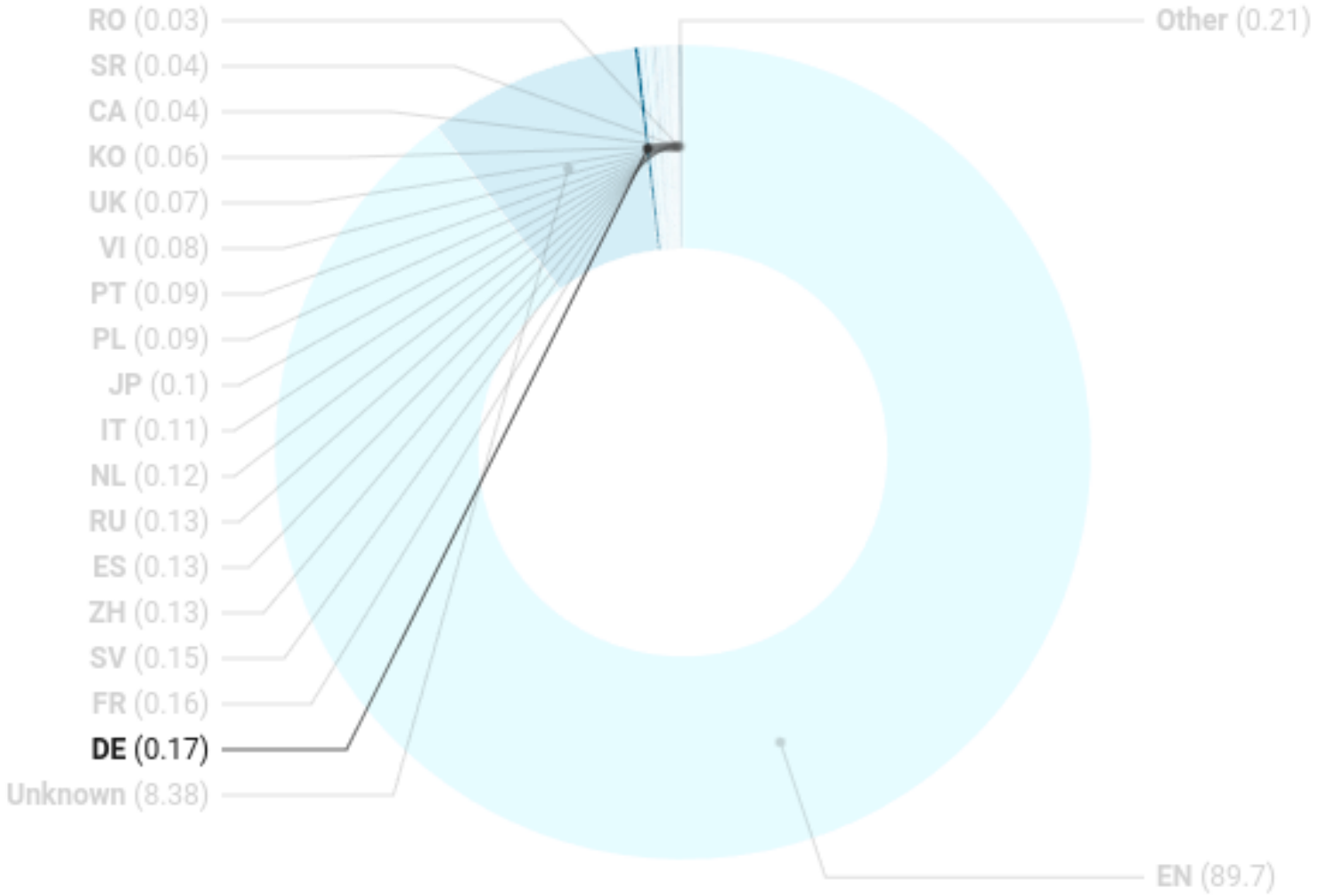
[Introducing DBRX: A New State-of-the-Art Open LLM | Databricks]

Llama 2's Pretrain-Data Language distribution



Llama 2's Language distribution in pretraining data with percentage | Created with Datawrapper

Llama 2's Pretrain-Data Language distribution

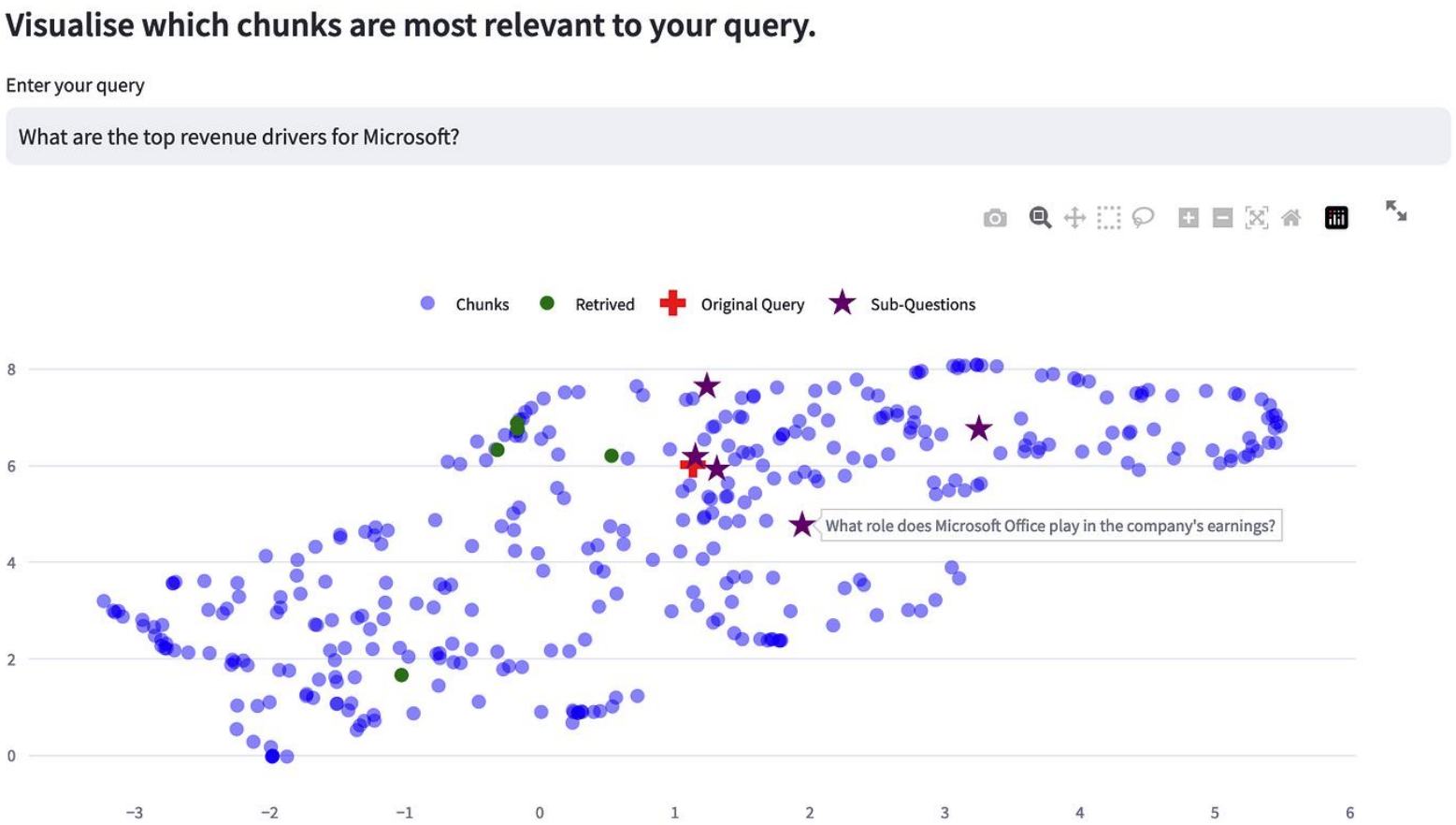


Llama 2's Language distribution in pretraining data with percentage | Created with Datawrapper

Debugging RAG

- Embedding
  - Hammock: colehaus/hammock-public: Visualize text embeddings (github.com)
  - RagExplorer - gabrielchua/RAGxplorer: Open-source tool to visualise your RAG 🔮 (github.com)

Debugging RAG

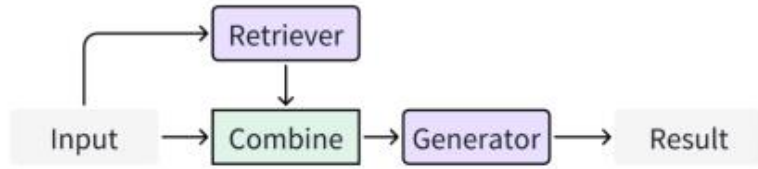


[gabrielchua/RAGxplorer: Open-source tool to visualise your RAG 🔮 (github.com)]

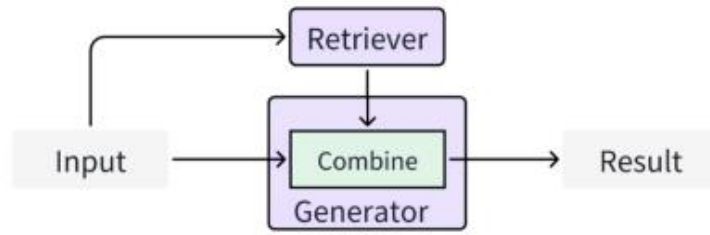

And this was only the tip of the iceberg...

RAG Foundations

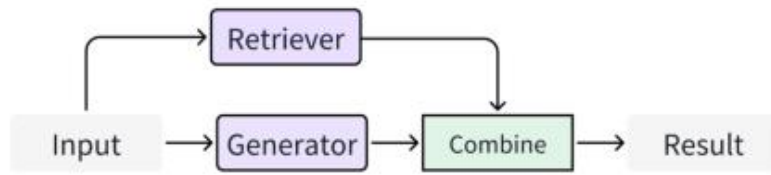
Query-based RAG
REALM, KILT, DocPrompting, etc.



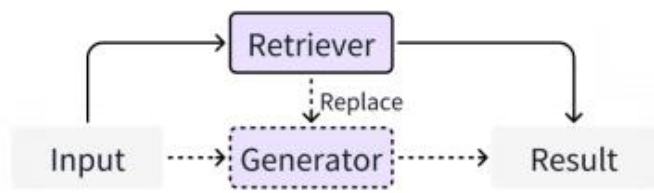
Latent Representation-based RAG
Retro, EaE, TOME, etc.

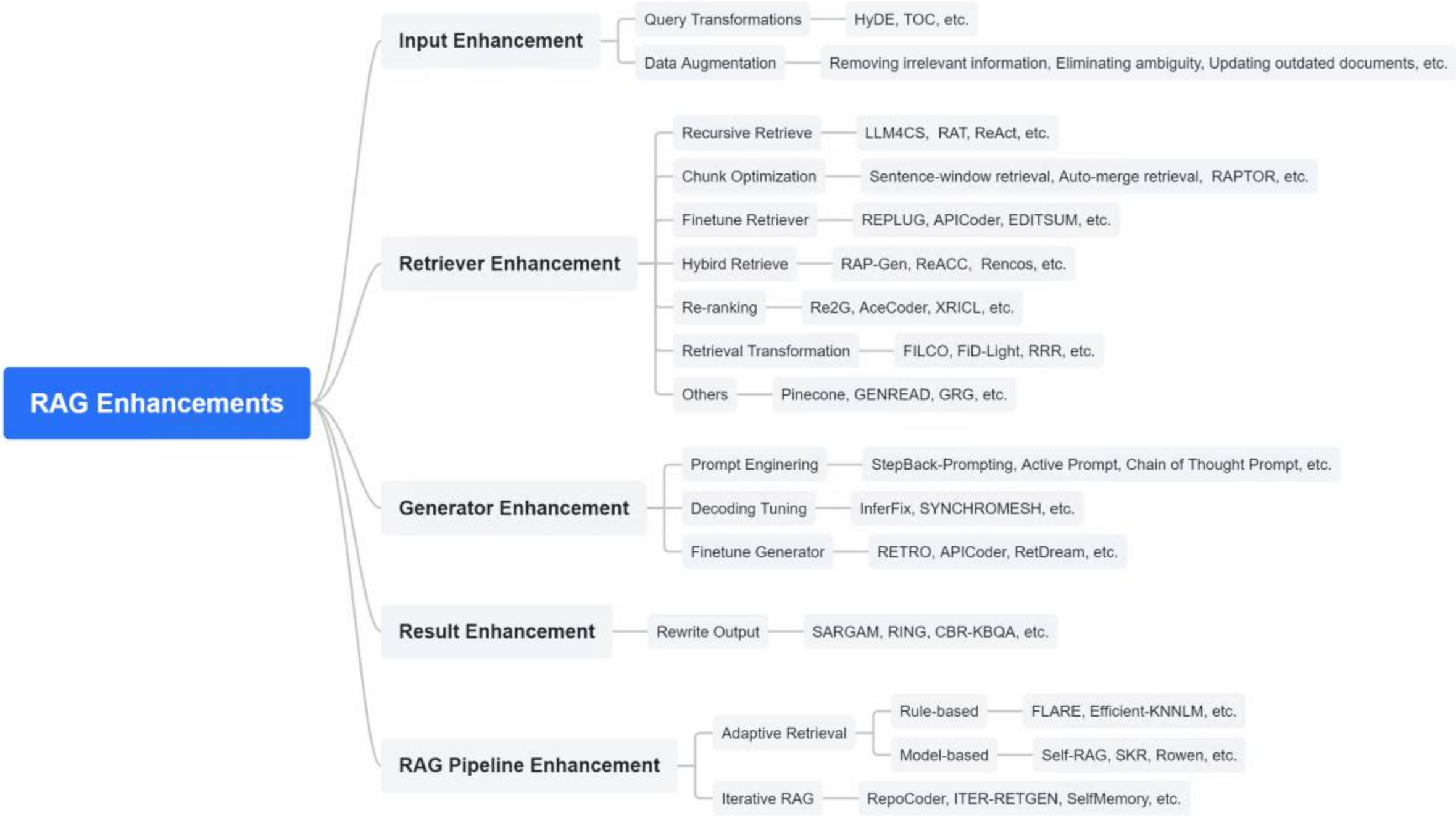


Logit-based RAG
kNN-LM, EDITSUM, TRIME, etc.



Speculative RAG
REST, GPTCache, etc.




RAG Enhancements
Input Enhancement
Query Transformations
HyDE, TOC, etc.
Data Augmentation
Removing irrelevant information, Eliminating ambiguity, Updating outdated documents, etc.
Retriever Enhancement
Recursive Retrieve
LLM4CS, RAT, ReAct, etc.
Chunk Optimization
Sentence-window retrieval, Auto-merge retrieval, RAPTOR, etc.
Finetune Retriever
REPLUG, APICoder, EDITSUM, etc.
Hybird Retrieve
RAP-Gen, ReACC, Rencos, etc.
Re-ranking
Re2G, AceCoder, XRICL, etc.
Retrieval Transformation
FILCO, FiD-Light, RRR, etc.
Others
Pinecone, GENREAD, GRG, etc.
Generator Enhancement
Prompt Enginering
StepBack-Prompting, Active Prompt, Chain of Thought Prompt, etc.
Decoding Tuning
InferFix, SYNCHROMESH, etc.
Finetune Generator
RETRO, APICoder, RetDream, etc.
Result Enhancement
Rewrite Output
SARGAM, RING, CBR-KBQA, etc.
RAG Pipeline Enhancement
Adaptive Retrieval
Rule-based
FLARE, Efficient-KNNLM, etc.
Model-based
Self-RAG, SKR, Rowen, etc.
Iterative RAG
RepoCoder, ITER-RETGEN, SelfMemory, etc.

Choices…

Indexing

- Data Preprocessing method
- Indexing Model
- Text-Splitting
- Chunking hyperparams

Storage

- Vector Database
- Adding Metadata

Retrieval

- Retrieval Method
- Top-K
- Similarity Cut-off

Synthesis and Generation

- Model choice
- Prompt
- Hyperparameters

Evaluation

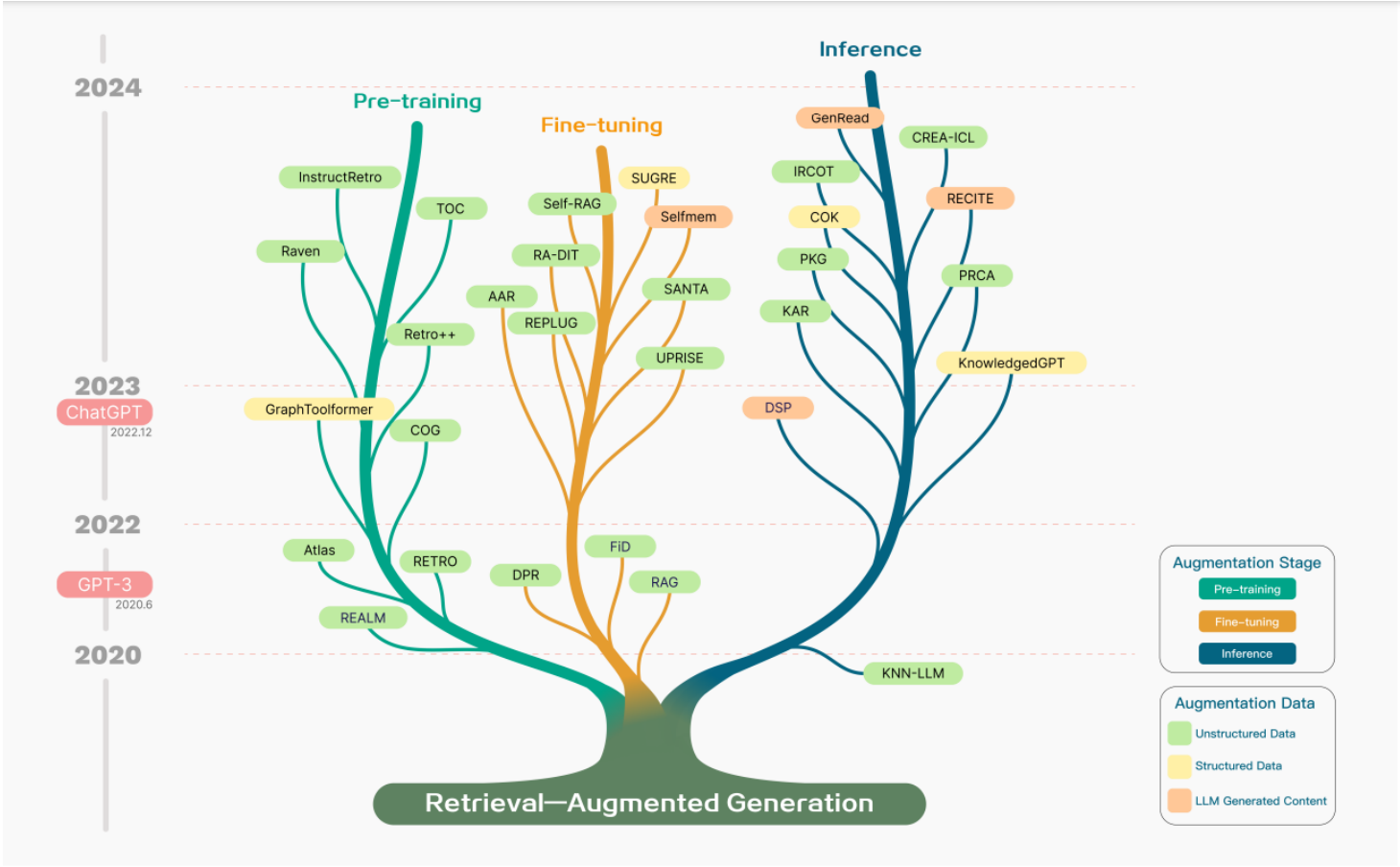
- Evaluation method
- Evaluator prompts

Take-Aways

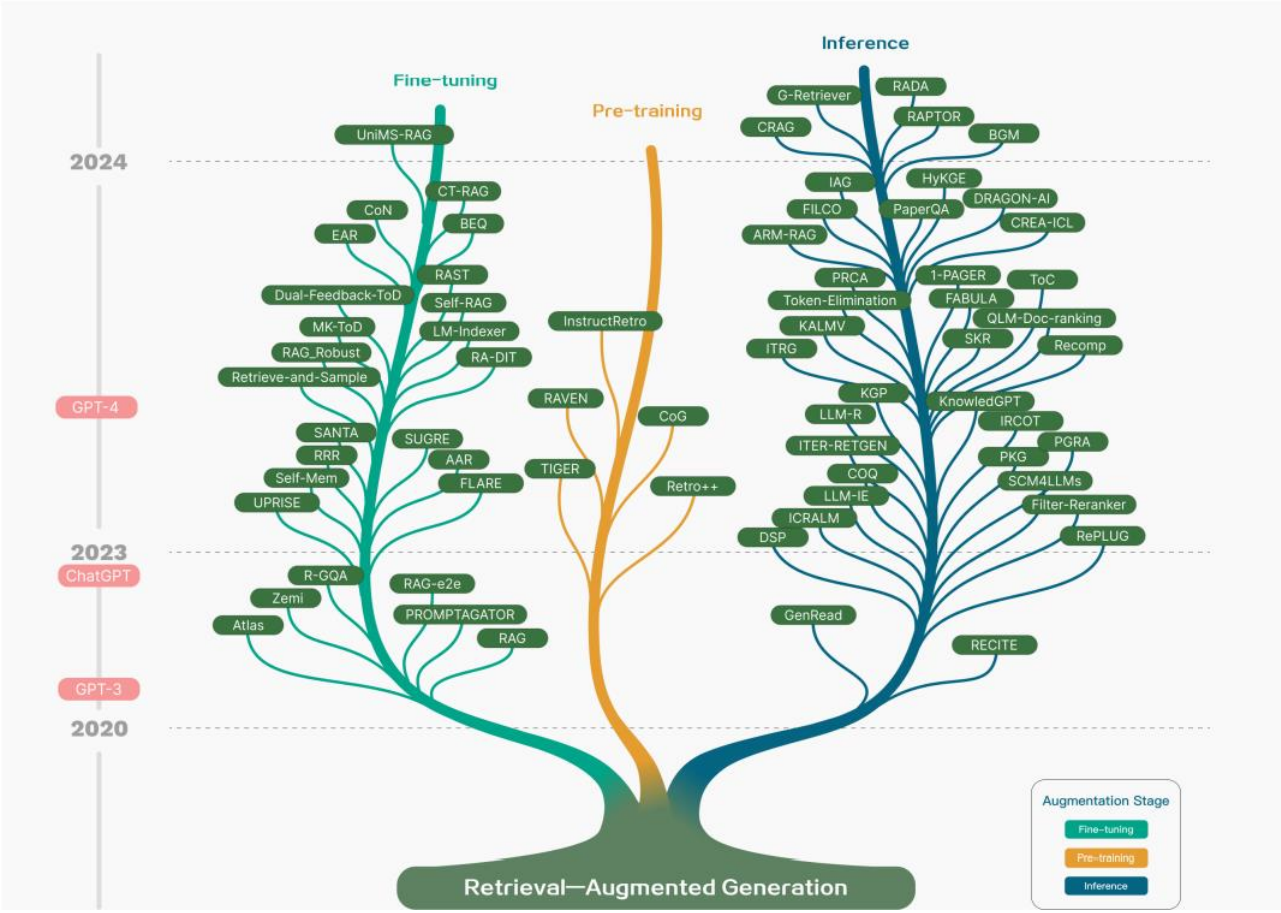
- RAG is a *user interface* for search
  - Search and fine the results through chat-bot LLMs
  - Hard to evaluate automatically (See Aaron Kaplan's talk)
- Not a single end-2-end model, but a complex of components
  - Hence, hard(er) to tune and reach high accuracy
- Huge community around it (and hype too)
  - Many new advances every week
- No one-solution-fits-all

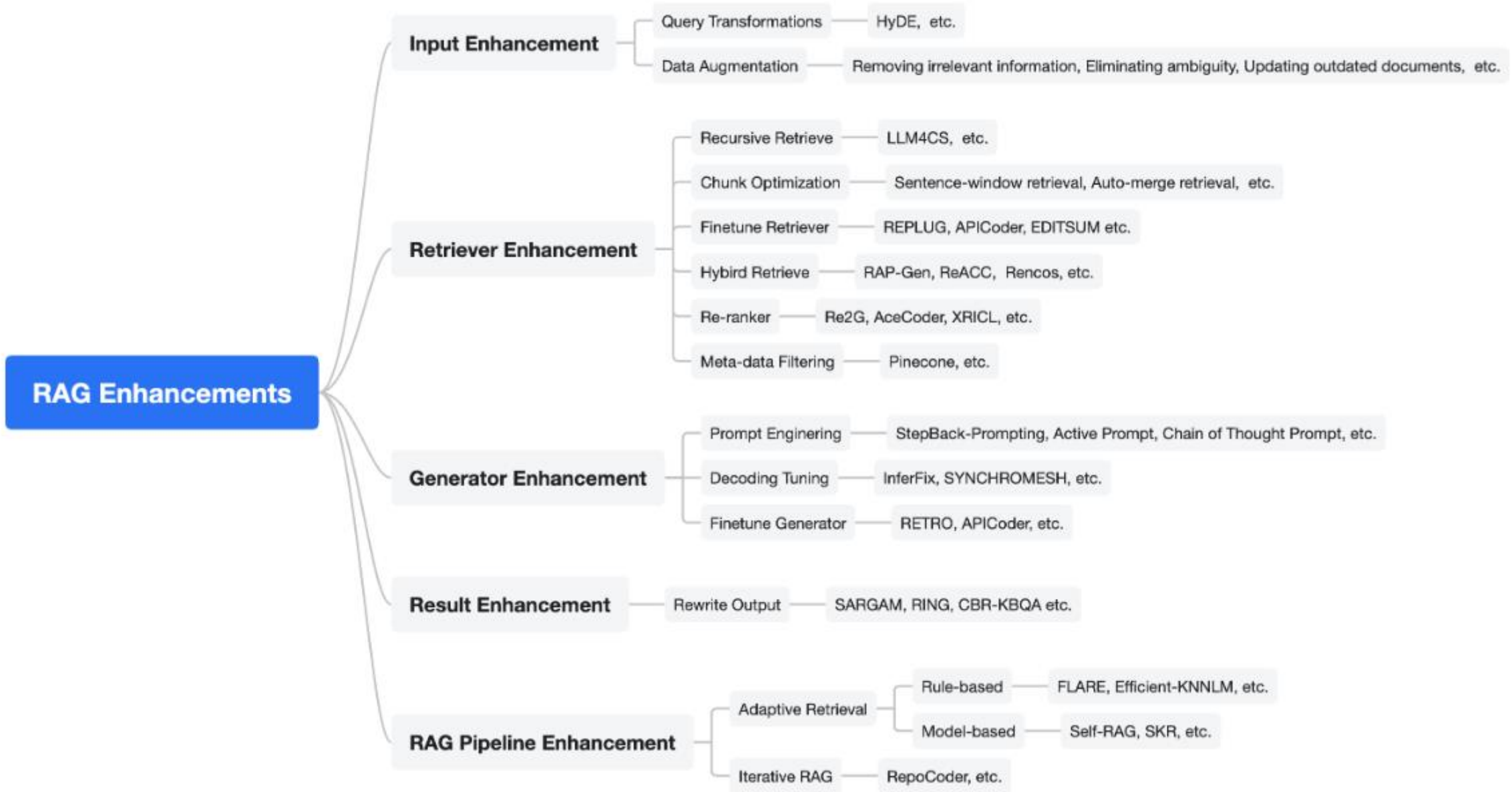

Thank you for your attention!

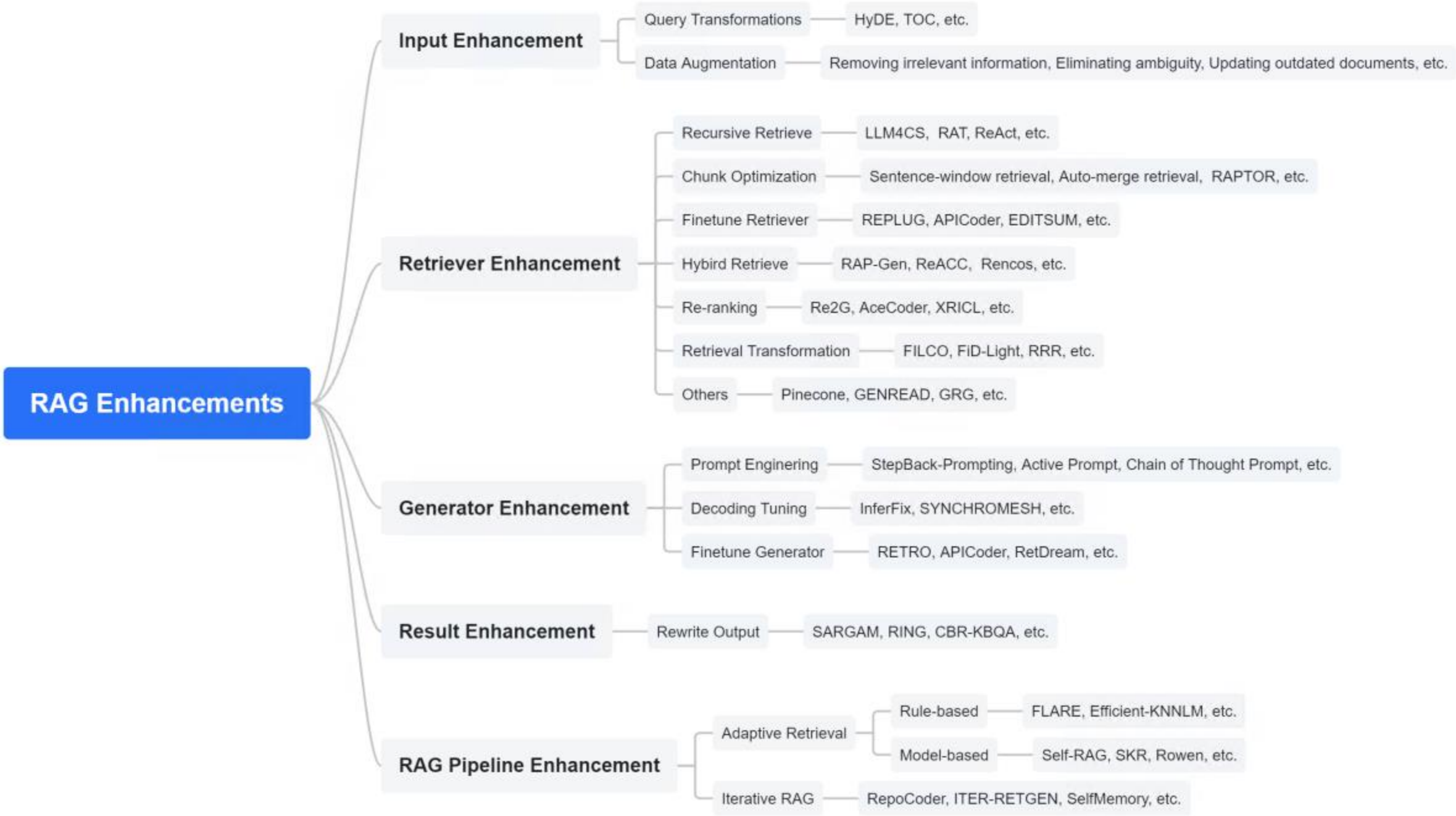
Supplementals – RAG Timeline (Jan 2024)



Supplementals – RAG Timeline (Apr 2024)




RAG Enhancements
Input Enhancement
Query Transformations
HyDE, etc.
Data Augmentation
Removing irrelevant information, Eliminating ambiguity, Updating outdated documents, etc.
Retriever Enhancement
Recursive Retrieve
LLM4CS, etc.
Chunk Optimization
Sentence-window retrieval, Auto-merge retrieval, etc.
Finetune Retriever
REPLUG, APICoder, EDITSUM etc.
Hybird Retrieve
RAP-Gen, ReACC, Rencos, etc.
Re-ranker
Re2G, AceCoder, XRICL, etc.
Meta-data Filtering
Pinecone, etc.
Generator Enhancement
Prompt Enginering
StepBack-Prompting, Active Prompt, Chain of Thought Prompt, etc.
Decoding Tuning
InferFix, SYNCHROMESH, etc.
Finetune Generator
RETRO, APICoder, etc.
Result Enhancement
Rewrite Output
SARGAM, RING, CBR-KBQA etc.
RAG Pipeline Enhancement
Adaptive Retrieval
Rule-based
FLARE, Efficient-KNNLM, etc.
Model-based
Self-RAG, SKR, etc.
Iterative RAG
RepoCoder, etc.


RAG Enhancements
Input Enhancement
Query Transformations
HyDE, TOC, etc.
Data Augmentation
Removing irrelevant information, Eliminating ambiguity, Updating outdated documents, etc.
Retriever Enhancement
Recursive Retrieve
LLM4CS, RAT, ReAct, etc.
Chunk Optimization
Sentence-window retrieval, Auto-merge retrieval, RAPTOR, etc.
Finetune Retriever
REPLUG, APICoder, EDITSUM, etc.
Hybird Retrieve
RAP-Gen, ReACC, Rencos, etc.
Re-ranking
Re2G, AceCoder, XRICL, etc.
Retrieval Transformation
FILCO, FiD-Light, RRR, etc.
Others
Pinecone, GENREAD, GRG, etc.
Generator Enhancement
Prompt Enginering
StepBack-Prompting, Active Prompt, Chain of Thought Prompt, etc.
Decoding Tuning
InferFix, SYNCHROMESH, etc.
Finetune Generator
RETRO, APICoder, RetDream, etc.
Result Enhancement
Rewrite Output
SARGAM, RING, CBR-KBQA, etc.
RAG Pipeline Enhancement
Adaptive Retrieval
Rule-based
FLARE, Efficient-KNNLM, etc.
Model-based
Self-RAG, SKR, Rowen, etc.
Iterative RAG
RepoCoder, ITER-RETGEN, SelfMemory, etc.

References

- [2312.10997v5] Retrieval-Augmented Generation for Large Language Models: A Survey (arxiv.org)
- [2312.10997] Retrieval-Augmented Generation for Large Language Models: A Survey (arxiv.org)
- [YouTube] Stanford CS25: V3 I Retrieval Augmented Language Models
- GitHub - Tongji-KGLLM/RAG-Survey
- [2002.08909] REALM: Retrieval-Augmented Language Model Pre-Training (arxiv.org)