When RAG Meets LFMs: Towards Retrieval-Augmented Large Foundation Models

Website: *https://advanced-rag.github.io/RAG-Meets-LFMs*

Survey: *https://arxiv.org/pdf/2405.06211*

Xu Yuan[1], Yujuan Ding[1], Chengliang Liu[1], Rui An[1], Chun-Hin Chan[1],
Yiqi Wang[2], Wenqi Fan[1], and Qing Li[1]

[1]The Hong Kong Polytechnic University

[2]National University of Defense Technology

June 9th (Day 1)

PAKDD 2026, Hong Kong, China

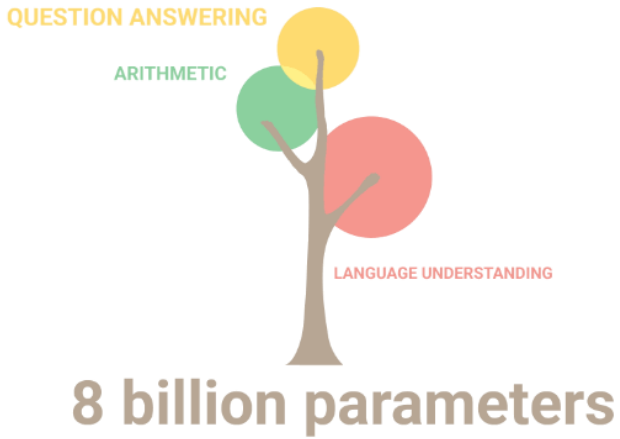
Tutorial Outline

- **Part 1: Introduction of Retrieval Augmented Large Foundation Models (RA-LFMs) (Dr. Wenqi Fan)**
- **Part 2: Architecture** of RA-LFMs and Main Modules (Xu Yuan)
- **Part 3: Learning** Approach of RA-LFMs (Chengliang Liu)
- **Part 4: Agentic RAG** (Chengliang Liu)
- **Part 5: Applications** of RA-LFMs (Chun-Hin Chan)
- **Part 6: Challenges and Future Directions** of RA-LFMs (Dr. Wenqi Fan)
- **Part 7: Q&A**

Website of this tutorial
Check out the slides and more information!

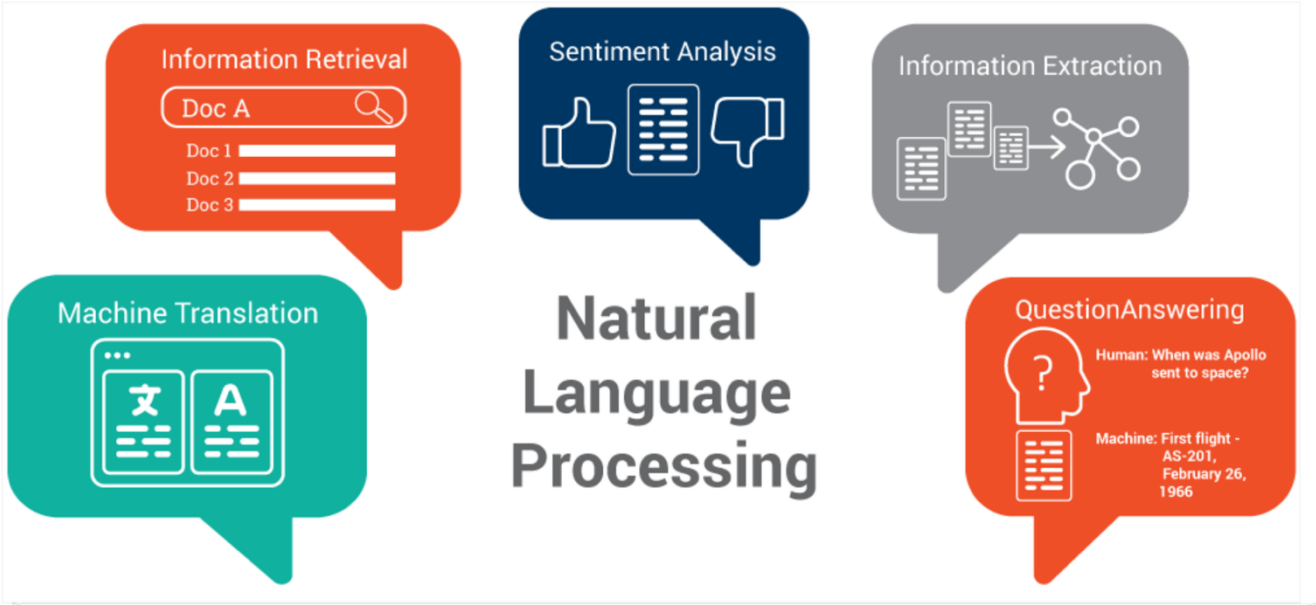
Large Foundation Models (LFMs)

OpenAI | Claude | Gemini | deepseek | .....

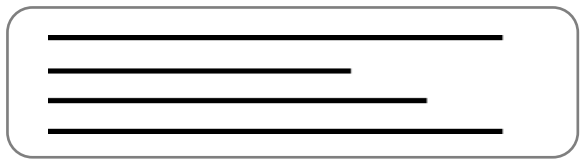


https://github.com/Hannibal046/Awesome-LLM/tree/main

Large Foundation Models (LFMs)



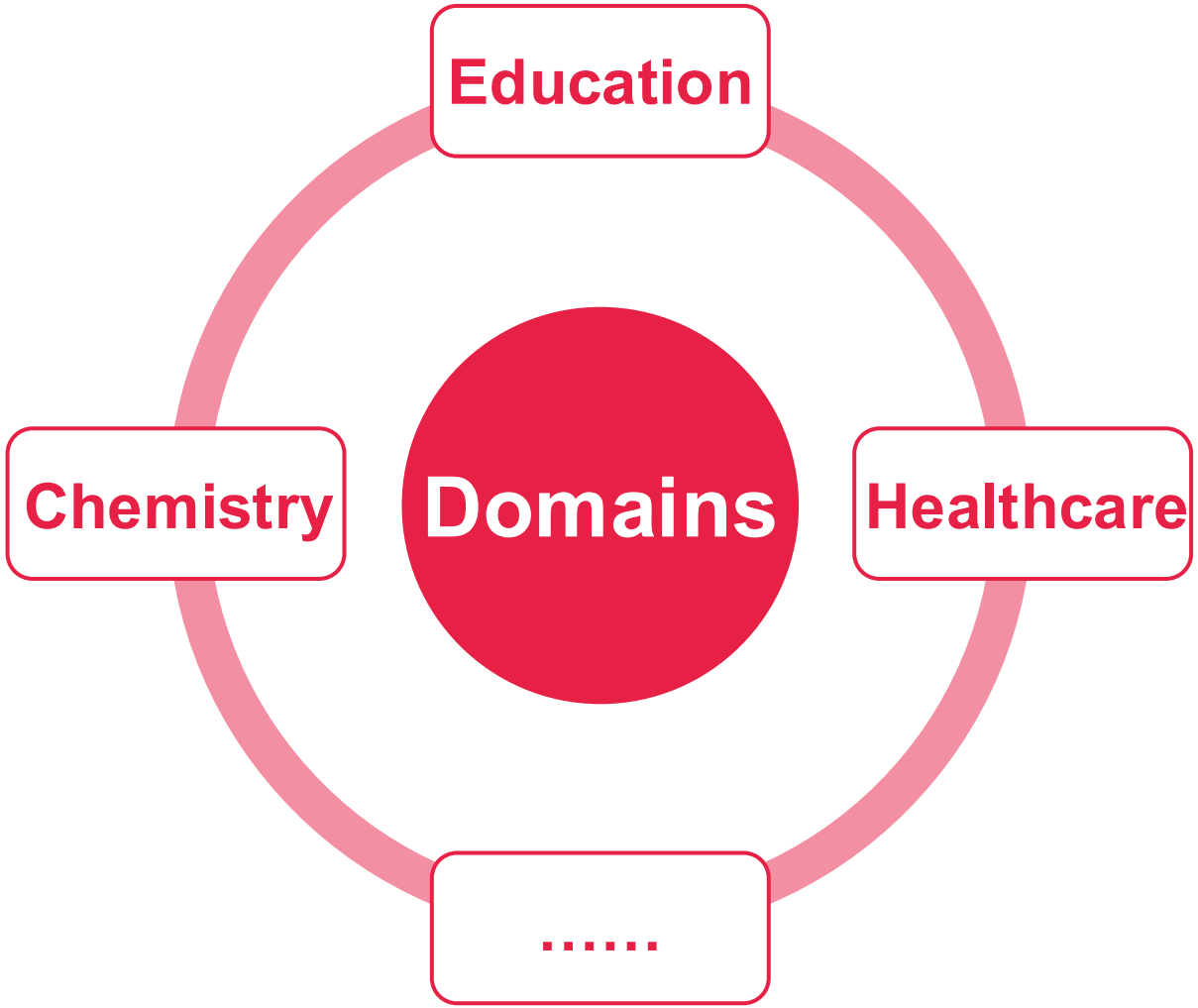
Input Text



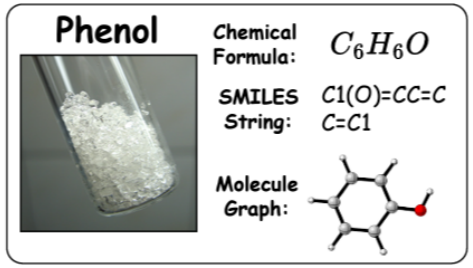
Generated Text

Large Foundation Models (LFMs)

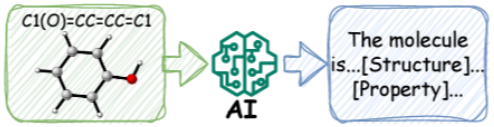
LFMs in Downstream Domains



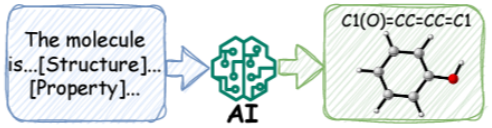
❑ **Molecule discovery**, etc.

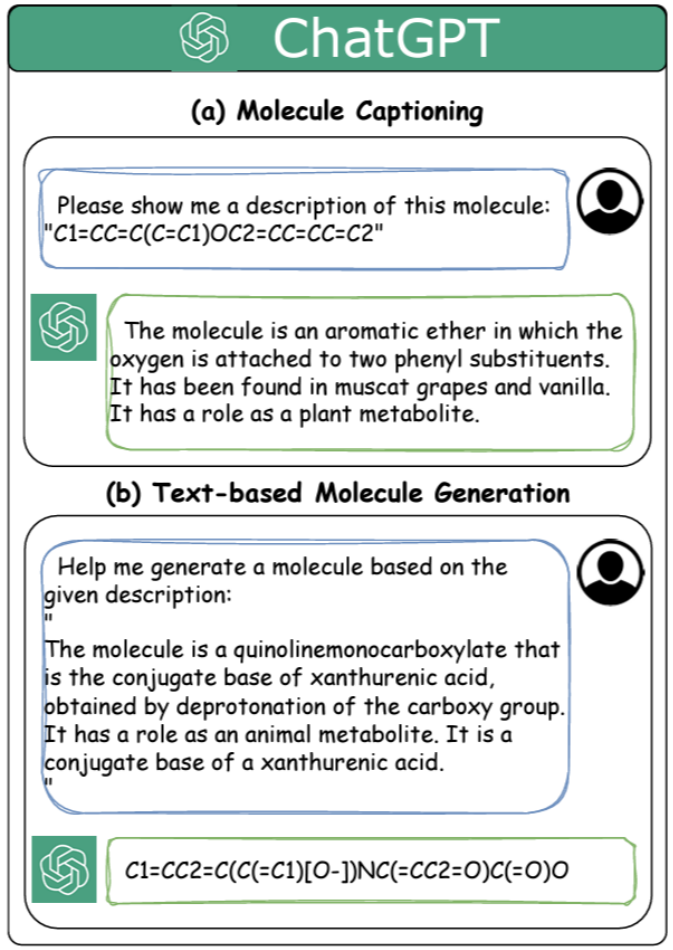


(a) Molecule Representations.



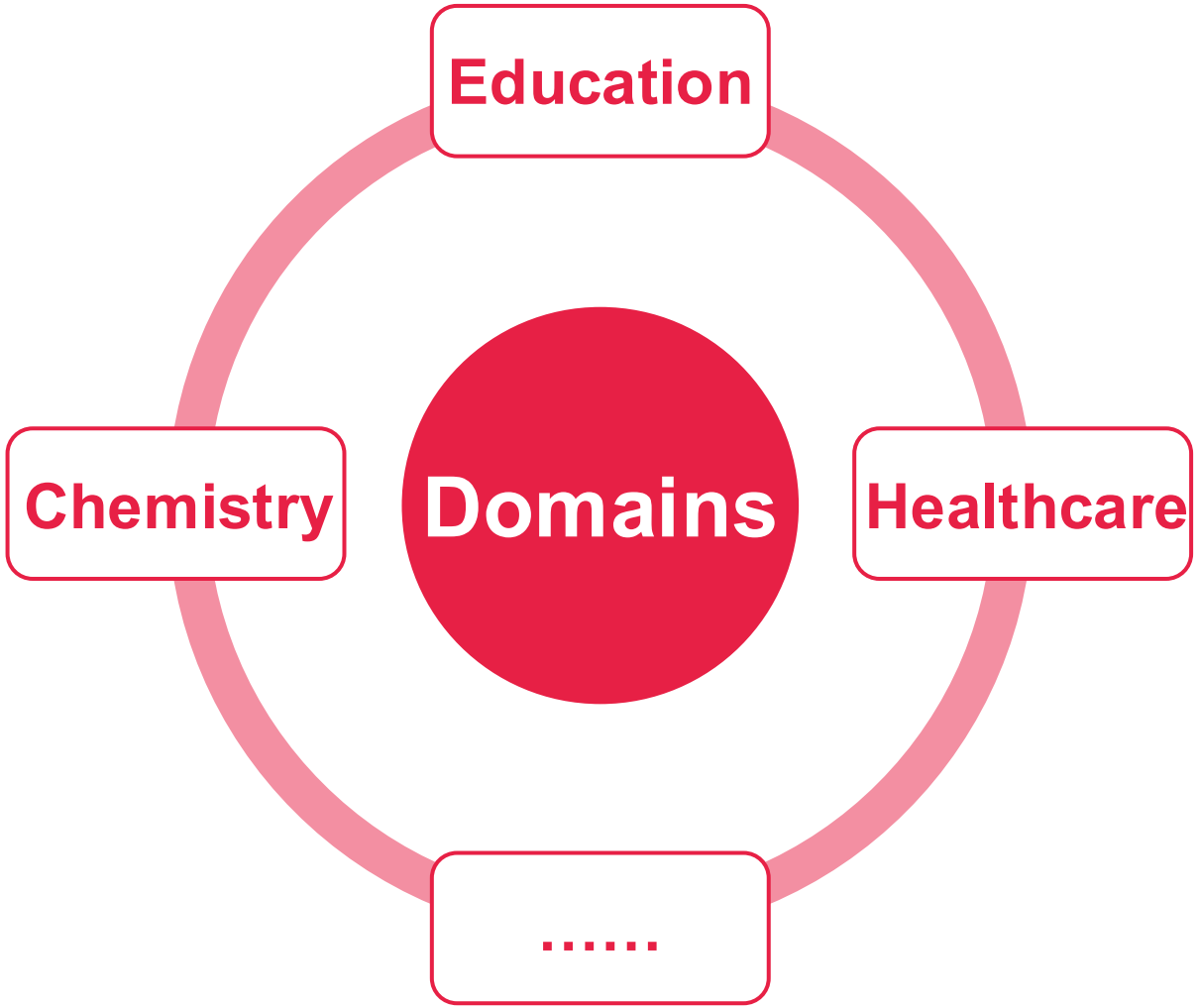
(b) Molecule Captioning.



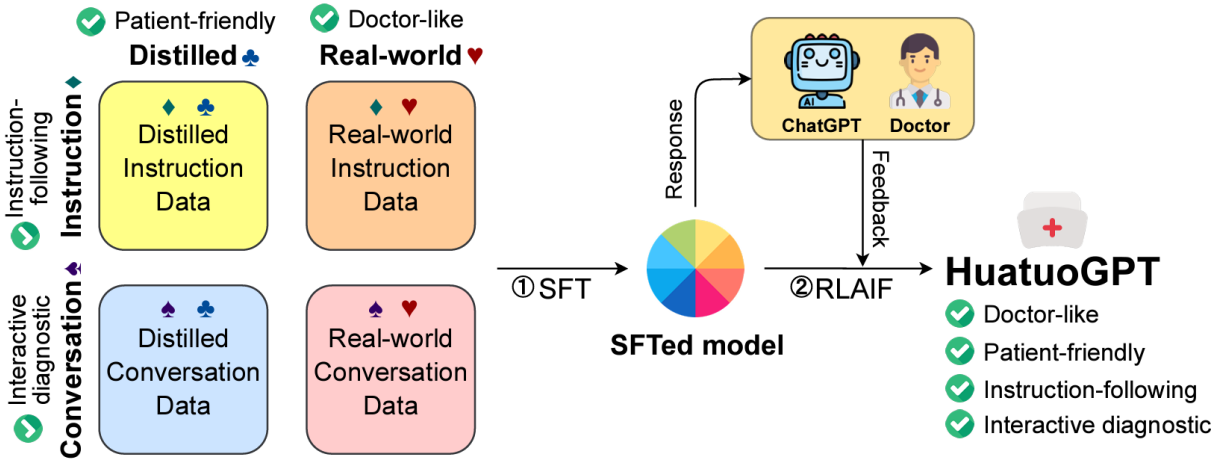


Li et al, 2024, Empowering Molecule Discovery for Molecule-Caption Translation with Large Language Models: A ChatGPT Perspective,
Liu et al., 2024, MolecularGPT: Open Large Language Model (LLM) for Few-Shot Molecular Property Prediction,

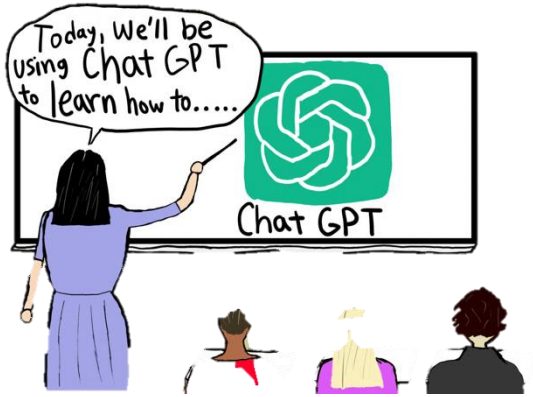
LFMs in Downstream Domains



❑ Medical consultation, etc.

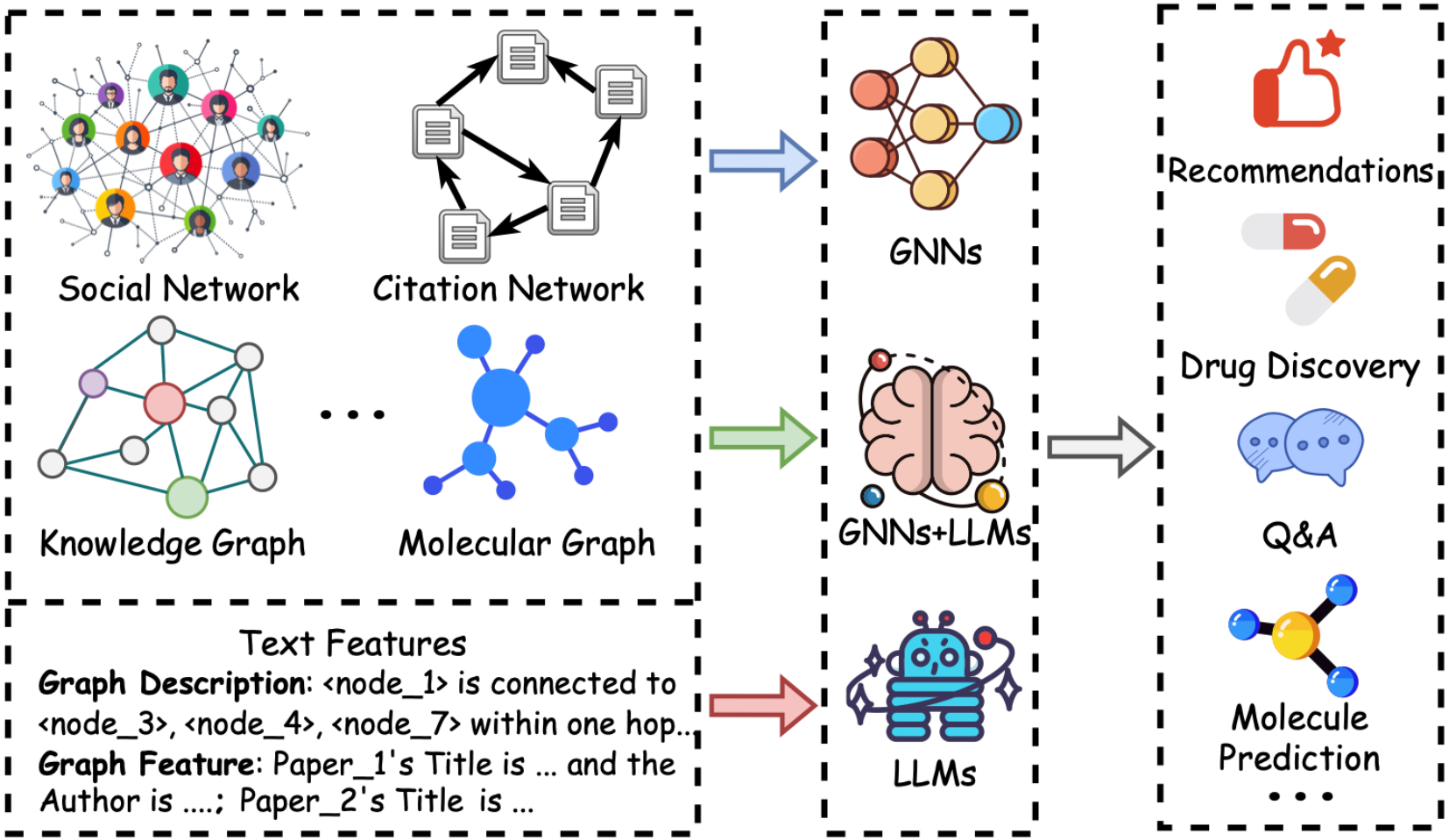


❑ Curriculum & Teaching, etc.



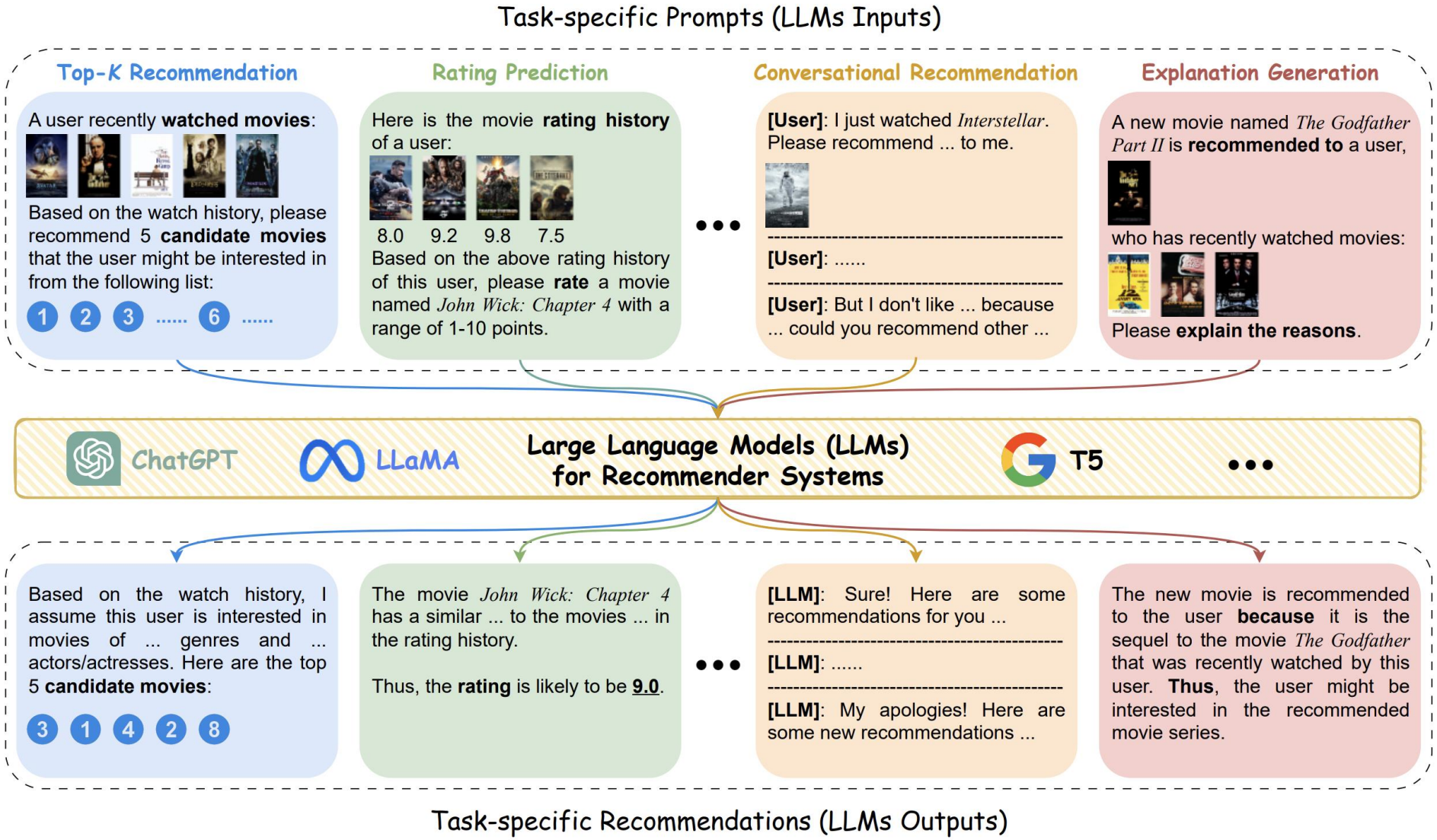
Zhang et al., 2023, HuatuoGPT, towards Taming Language Model to Be a Doctor

LFMs on Graph-structured Data



Fan, et al., 2024, Graph Machine Learning in the Era of Large Language Models (LLMs).

LFMs in Recommender Systems



Zhao, et al., 2024, Recommender systems in the era of large language models (llms).

Challenges and Risks of LFMs

☐ Outdated Parametric Knowledge

LFMs rely on knowledge stored in model parameters, which may become outdated and cannot easily reflect newly updated documents, policies, or domain facts.

☐ Ungrounded Hallucination

Without access to external evidence, LFMs may generate plausible but unsupported answers, especially for knowledge-intensive or factual queries.

☐ Domain Knowledge Gap

LFMs often lack fine-grained expertise in specialized domains, making them less reliable for tasks requiring professional or private-domain knowledge.

☐ Limited Traceability & Verifiability

LFMs usually do not provide explicit sources for their responses, making it difficult to verify, audit, or trust the generated content.

LFMs' Challenges in Vertical Domains

- Domain of Law

OXFORD

Journal of Legal Analysis, 2024, **16**, 64–93

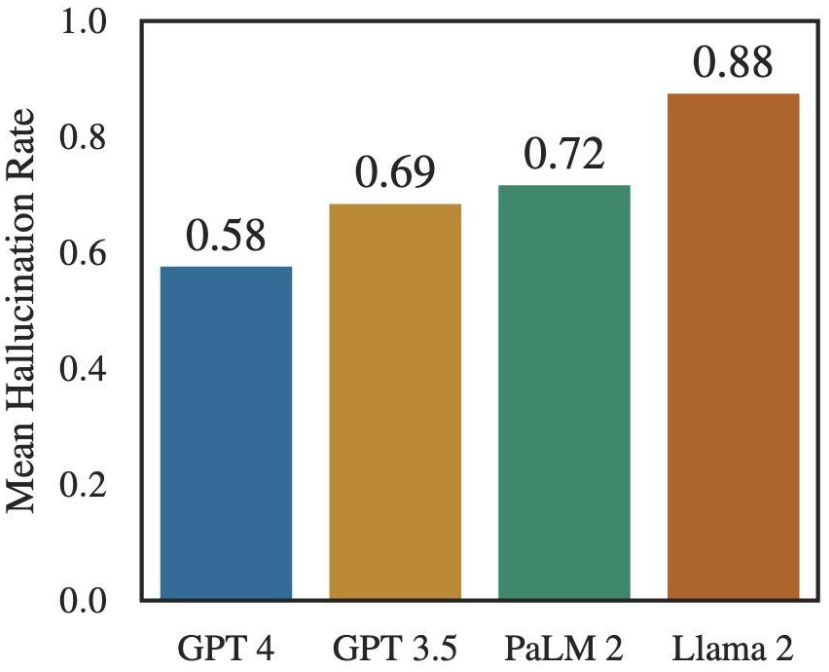
https://doi.org/10.1093/jla/laae003
Advance access publication 26 June 2024
Article

Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models

Matthew Dahl*, Varun Magesh†, Mirac Suzgun‡, and Daniel E. Ho§

*In a new study by **Stanford RegLab** and **Institute for Human-Centered AI** researchers, it is demonstrated that legal hallucinations are pervasive and disturbing: **hallucination rates range from 69% to 88% in response to specific legal queries** for state-of-the-art language models.*

Hallucinations are common across all LFMs when they are asked a direct, verifiable question about a federal court case

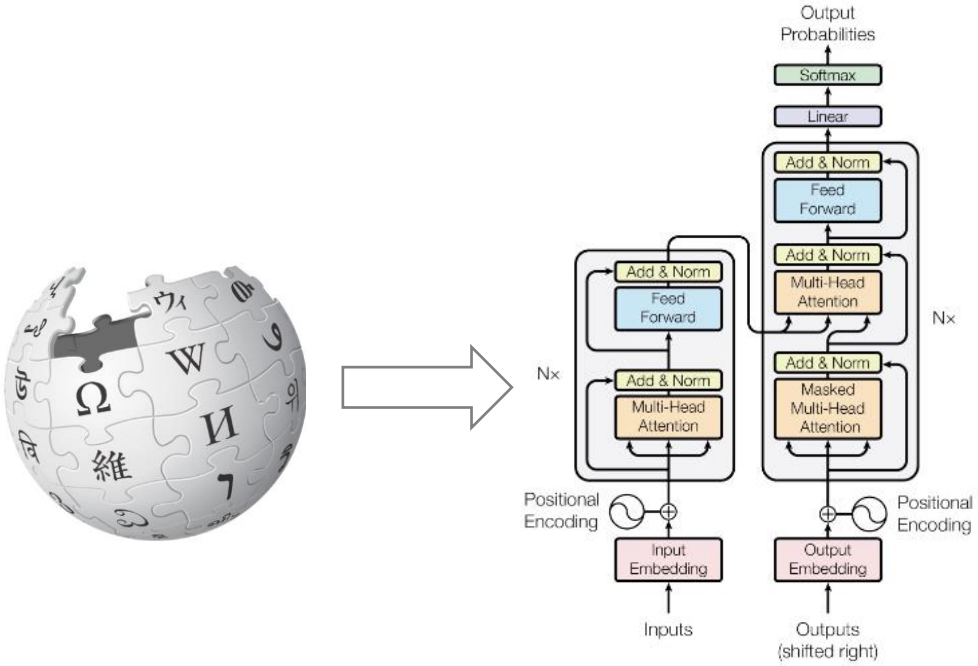


Dahl M, et al. 2024, Large legal fictions: Profiling legal hallucinations in large language models.

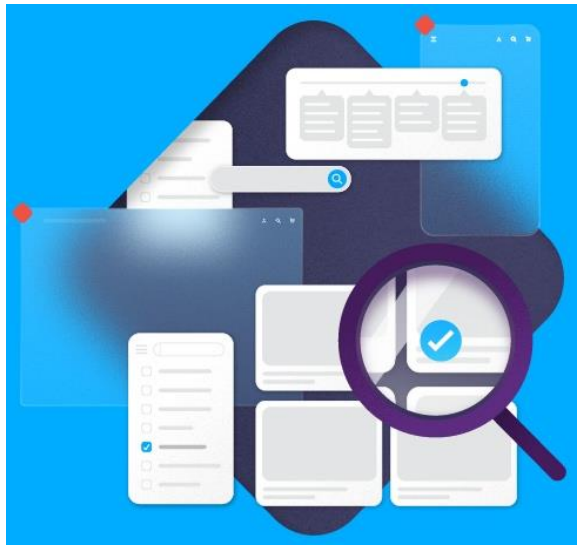
Why Large Foundation Models Work Well?

- Big Model + Big Training Data

Storing knowledge in the parametric model !



Storing knowledge in the non-parametric model?

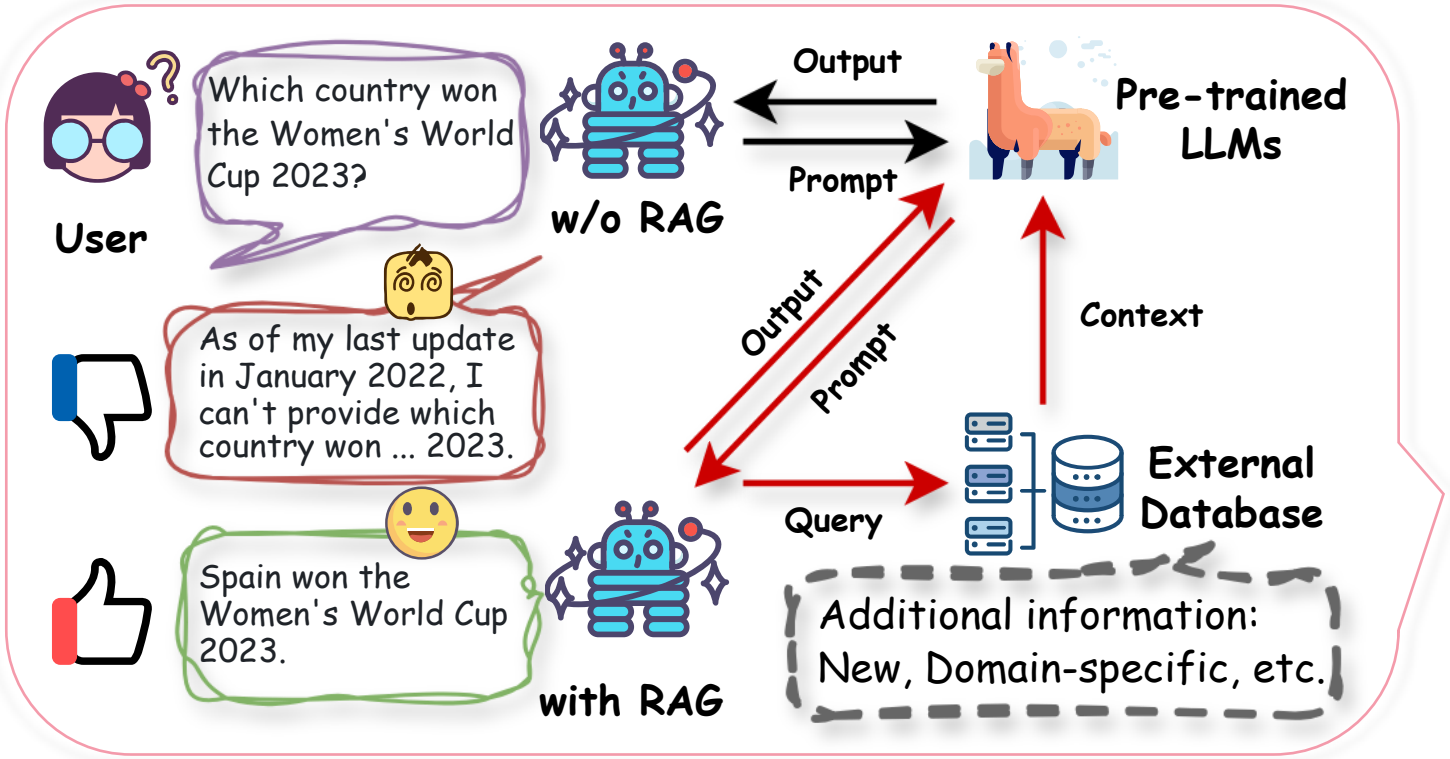


Information Retrieval (IR)

Retrieval-Augmented Large Foundation Models (RA-LFMs)

- LFMs **cannot memorize all** (particularly long-tail) knowledge in their parameters
- Lack of **domain-specific knowledge**, **updated information**, etc

Hallucination & Unable to answer → Re-training / Finetuning ?



Costly & Heavy Work

Retrieval-Augmented Generation (RAG) for LFM: **RA-LFMs**

Integrating Information Retrieval in Generation: **RA-LFM**

Data for Training LFMs
- Low quality
- General
- Fixed
- Hard to update

External Knowledge Base
- High-quality knowledge
- Specialized knowledge
- Scalable
- Easy-updated

+

=

Content generation

Close-book exam

*(Hard mode, have to **remember everything**)*

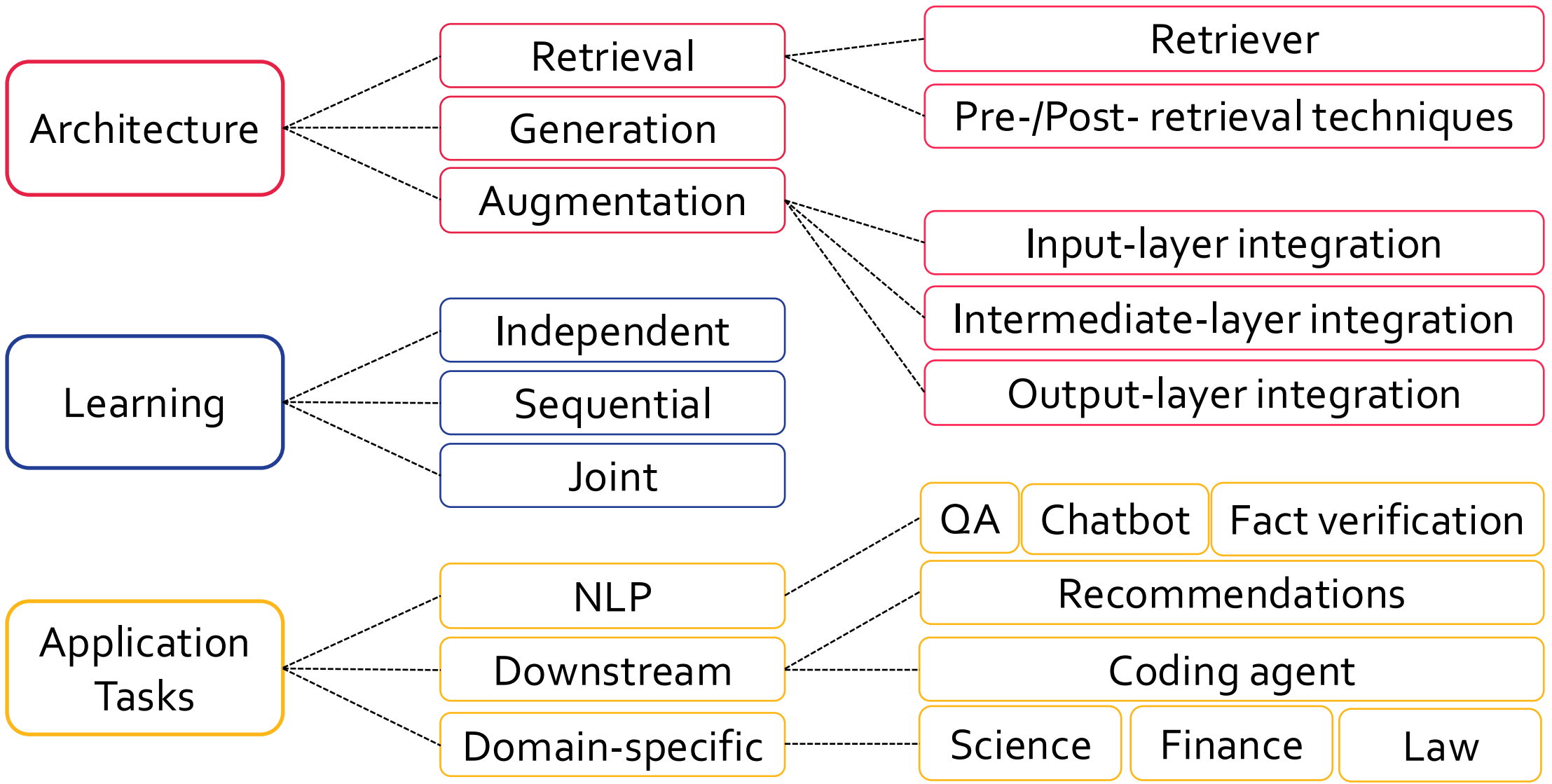
Information / Knowledge retrieval

RA-LFMs

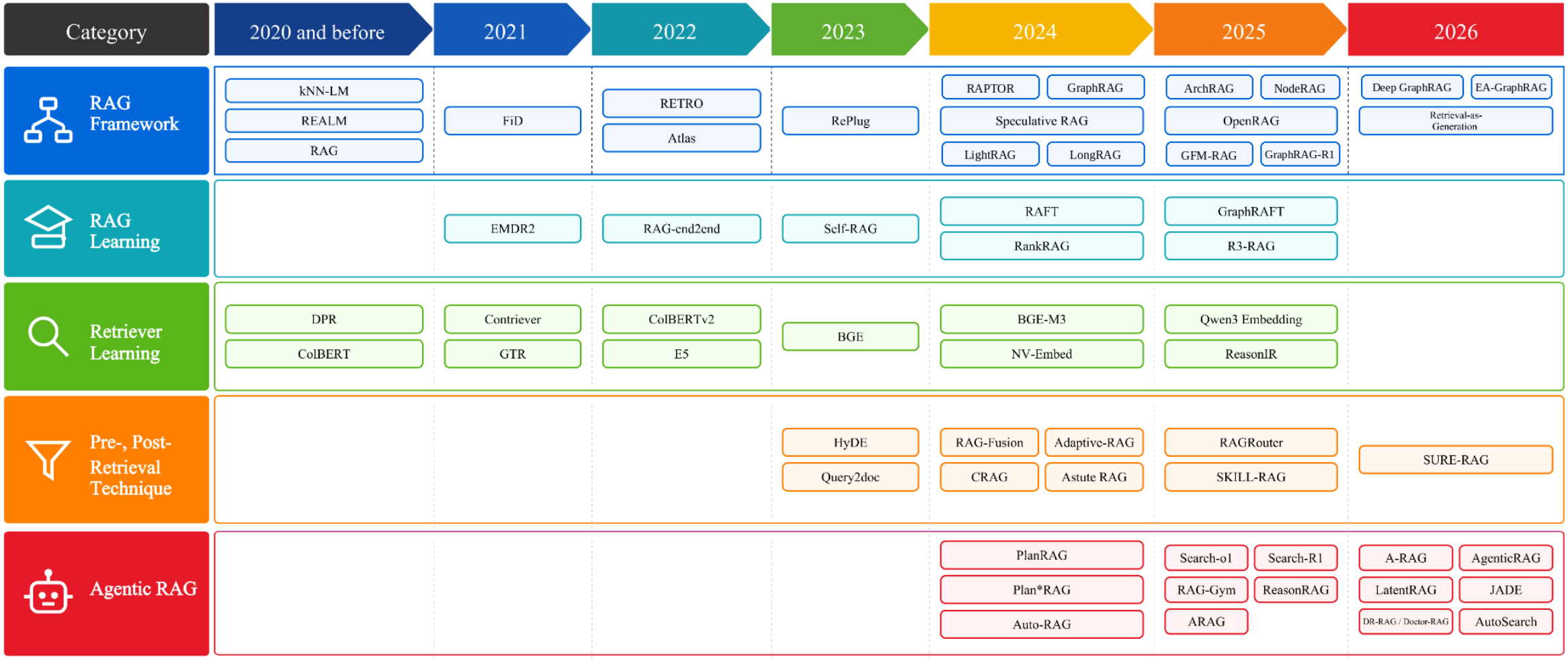
Open-book exam

(Easy mode, allow to search in reference)

RA-LFM Research Taxonomy



RAG & RA-LFM Model Development



A Comprehensive Survey Paper

A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models

Wenqi Fan
wenqifan03@gmail.com
The Hong Kong Polytechnic University, HK SAR

Yujuan Ding*
dingyujuan385@gmail.com
The Hong Kong Polytechnic University, HK SAR

Liangbo Ning
BigLemon1123@gmail.com
The Hong Kong Polytechnic University, HK SAR

Shijie Wang
shijie.wang@connect.polyu.hk
The Hong Kong Polytechnic University, HK SAR

Hengyun Li
neilhengyun.li@polyu.edu.hk
The Hong Kong Polytechnic University, HK SAR

Dawei Yin
yindawei@acm.org
Baidu Inc, China

Tat-Seng Chua
dcscts@nus.edu.sg
National University of Singapore, Singapore

Qing Li
csqli@comp.polyu.edu.hk
The Hong Kong Polytechnic University, HK SAR

Website of this tutorial
Check out the slides and more information!

Tutorial Outline

- **Part 1: Introduction** of Retrieval Augmented Large Foundation Models (RA-LFMs) (Dr. Wenqi Fan)
- **Part 2: Architecture of RA-LFMs and Main Modules (Xu Yuan)**
- **Part 3: Learning** Approach of RA-LFMs (Chengliang Liu)
- **Part 4: Agentic RAG** (Chengliang Liu)
- **Part 5: Applications** of RA-LFMs (Chun-Hin Chan)
- **Part 6: Challenges and Future Directions** of RA-LFMs (Dr. Wenqi Fan)
- **Part 7: Q&A**

Website of this tutorial
Check out the slides and more information!

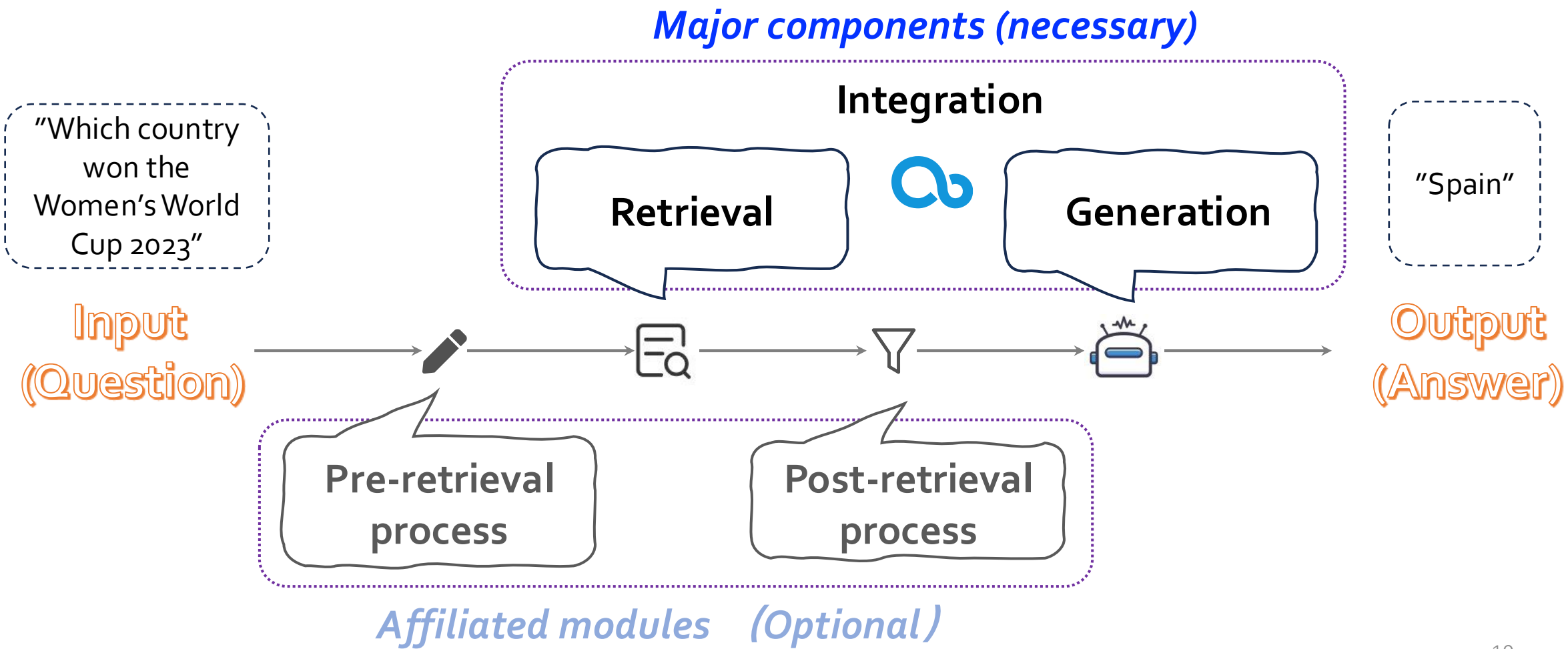
PART 2: Architecture of RA-LFMs and Main Modules

Presenter
Xu Yuan
HK PolyU

- **RA-LFM architecture overview**
- **Retriever in RA-LFMs**
- **Retrieval results integration**
- **Pre/Post-retrieval techniques**
- **Special RA-LFM paradigms**

RA-LFM Architecture: Standard Pipeline

- **Technical component illustration in a RA-LFM for the Q&A task**



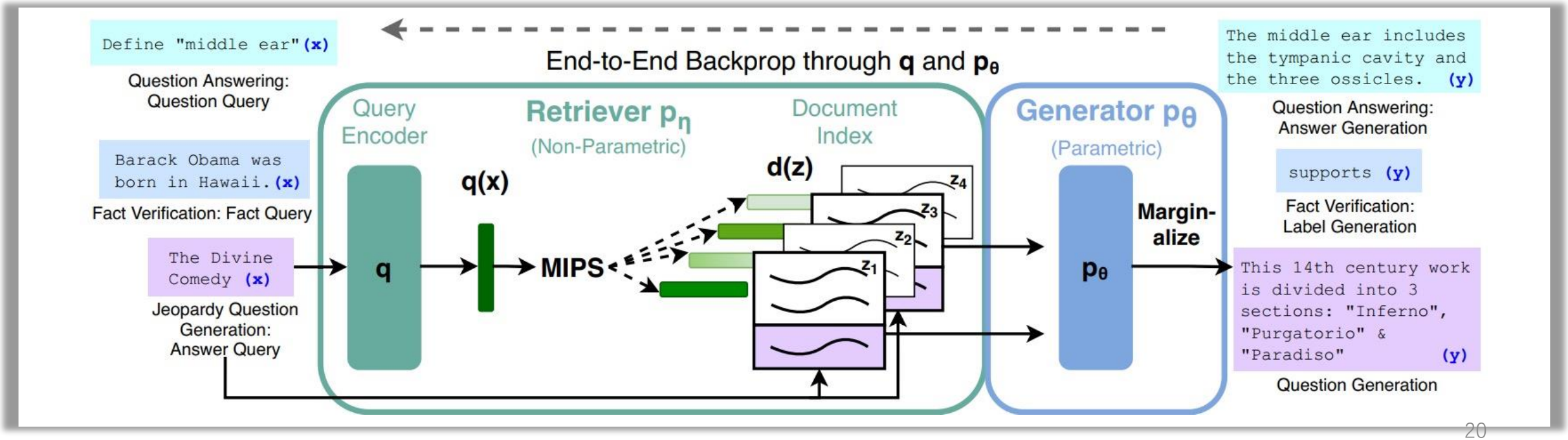
A Simple Retrieval-Augmented Generation Model

- **RAG**

Integration: concatenating each retrieved passage with the question

Retrieval: DPR + MIPS

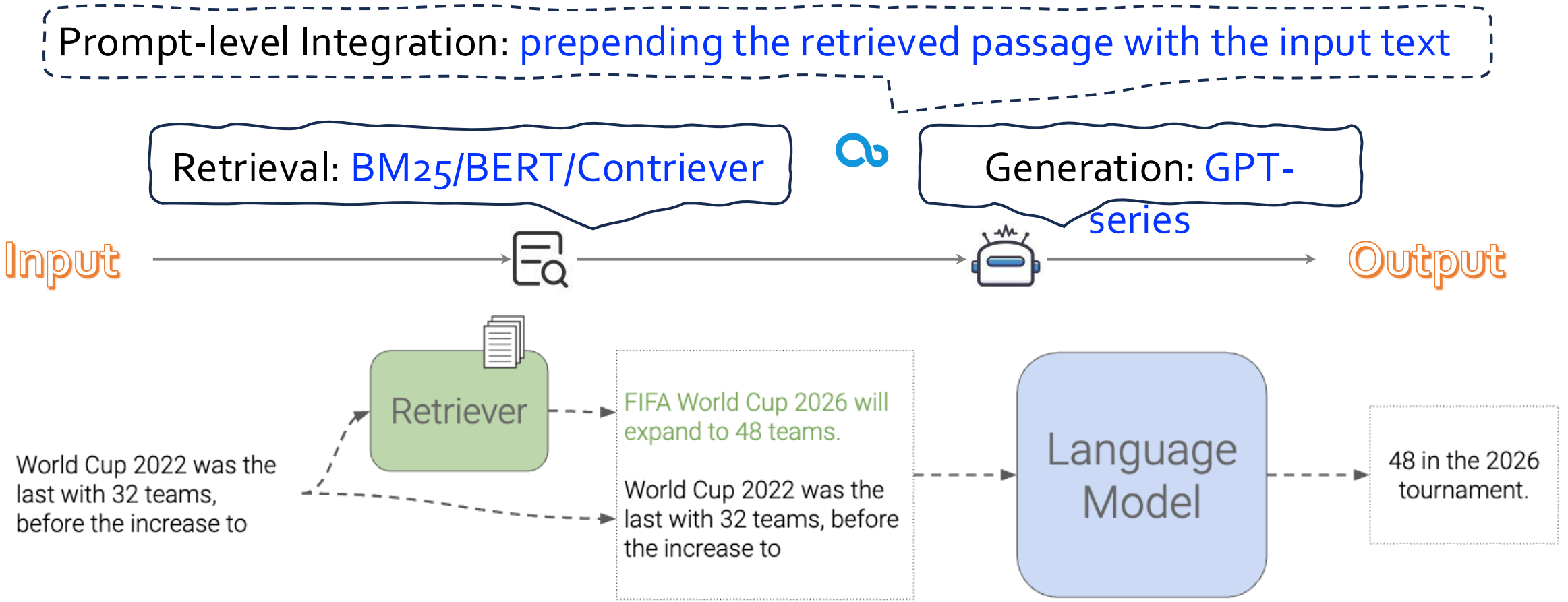
Generation: BART



Lewis et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." NeurIPS 2020

A Simple Retrieval-Augmented Generation Model

- **In-Context RALM**



Ram et al. 2023, "In-Context Retrieval-Augmented Language Models." TACL 2023

PART 2: Architecture of RA-LFMs and Main Modules


Slides

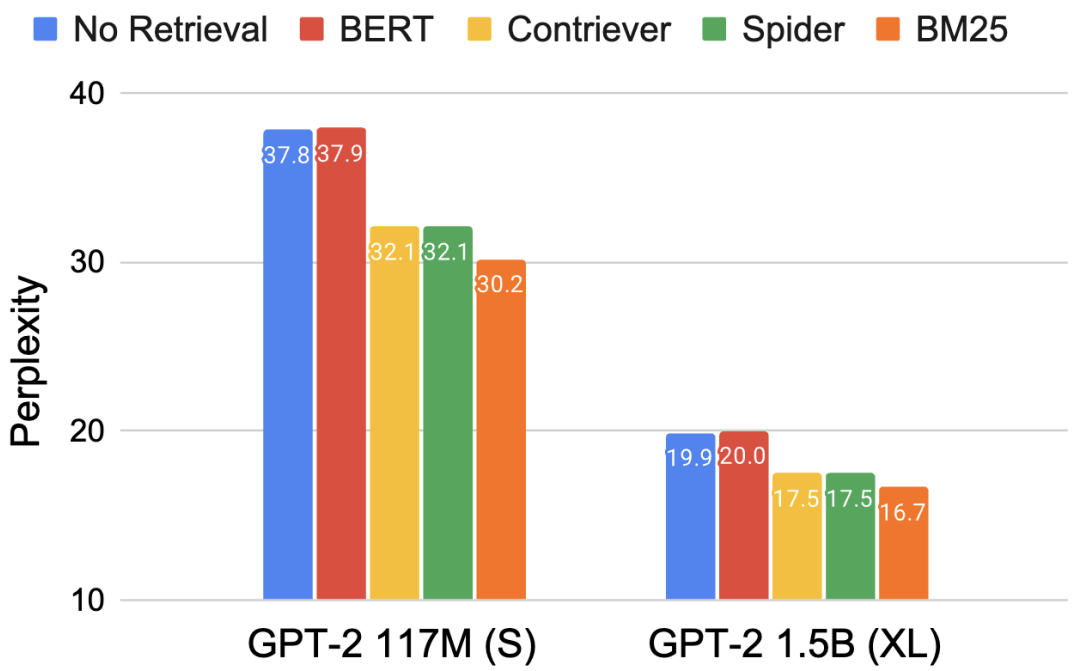
Website of this tutorial

- RA-LFM architecture overview
- Retriever in RA-LFMs
- Retrieval results integration
- Pre/Post-retrieval techniques
- Special RA-LFM paradigms

RA-LFM Architecture: Retriever Types

- Different types of retriever deliver different generation performance

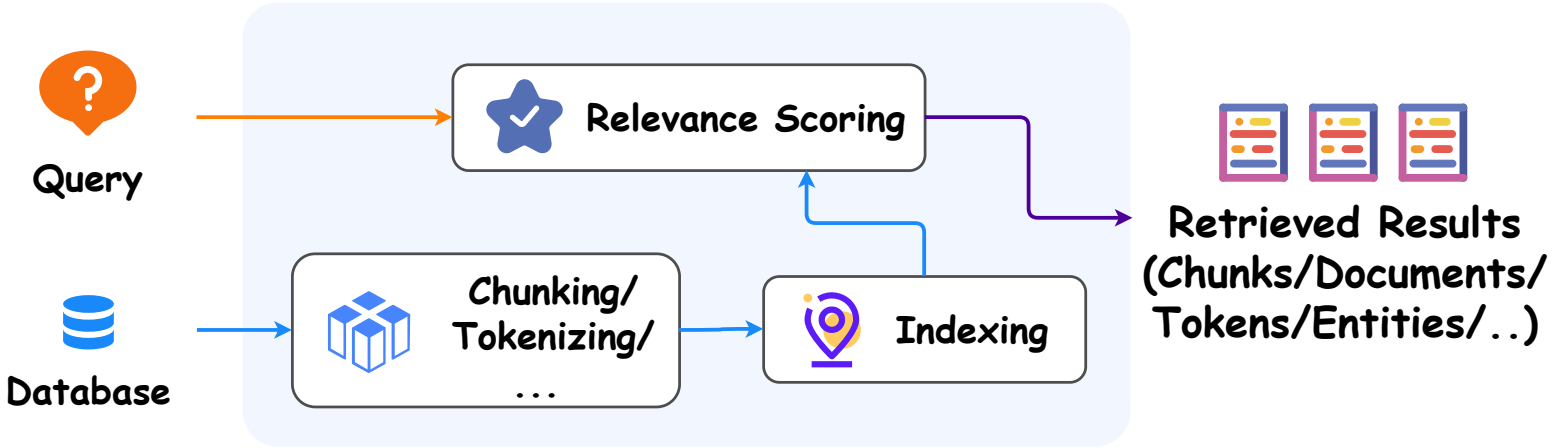
Relevance measurement	Retriever learning
Sparse	Task-specific Learning
Dense	General-purpose Learning



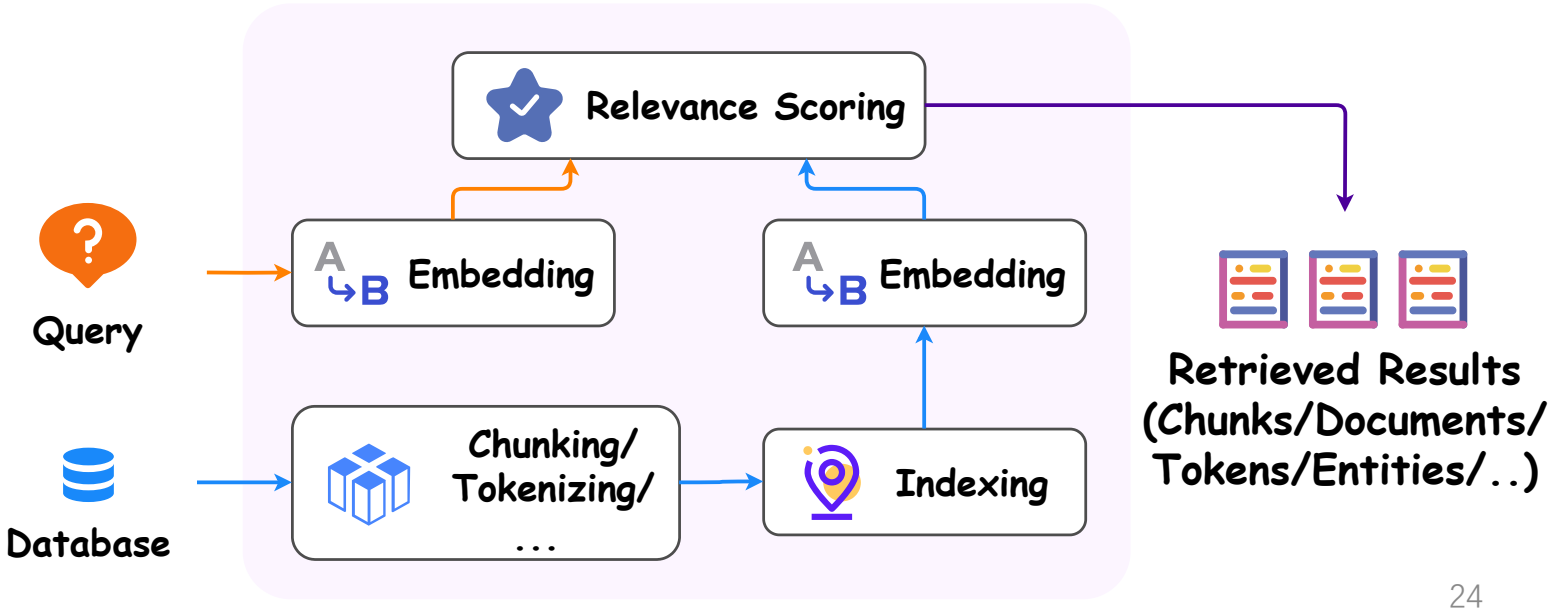
Ram et al. 2023, "In-Context Retrieval-Augmented Language Models." TACL 2023

Dense v.s. Sparse Retrievers

Sparse Retrievers (SR)



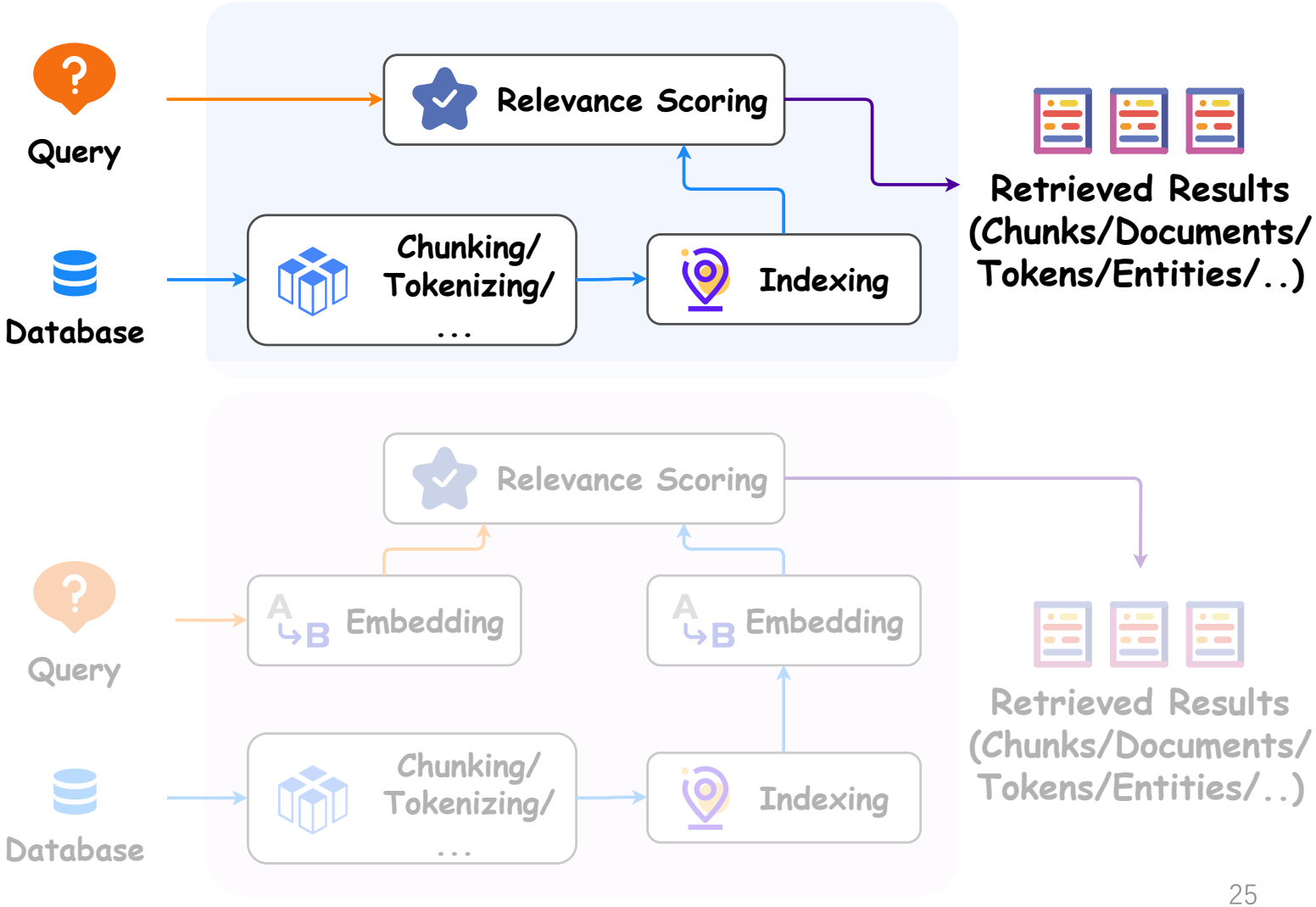
Dense Retrievers (DR)



Dense v.s. Sparse Retrievers

Sparse Retrievers (SR)

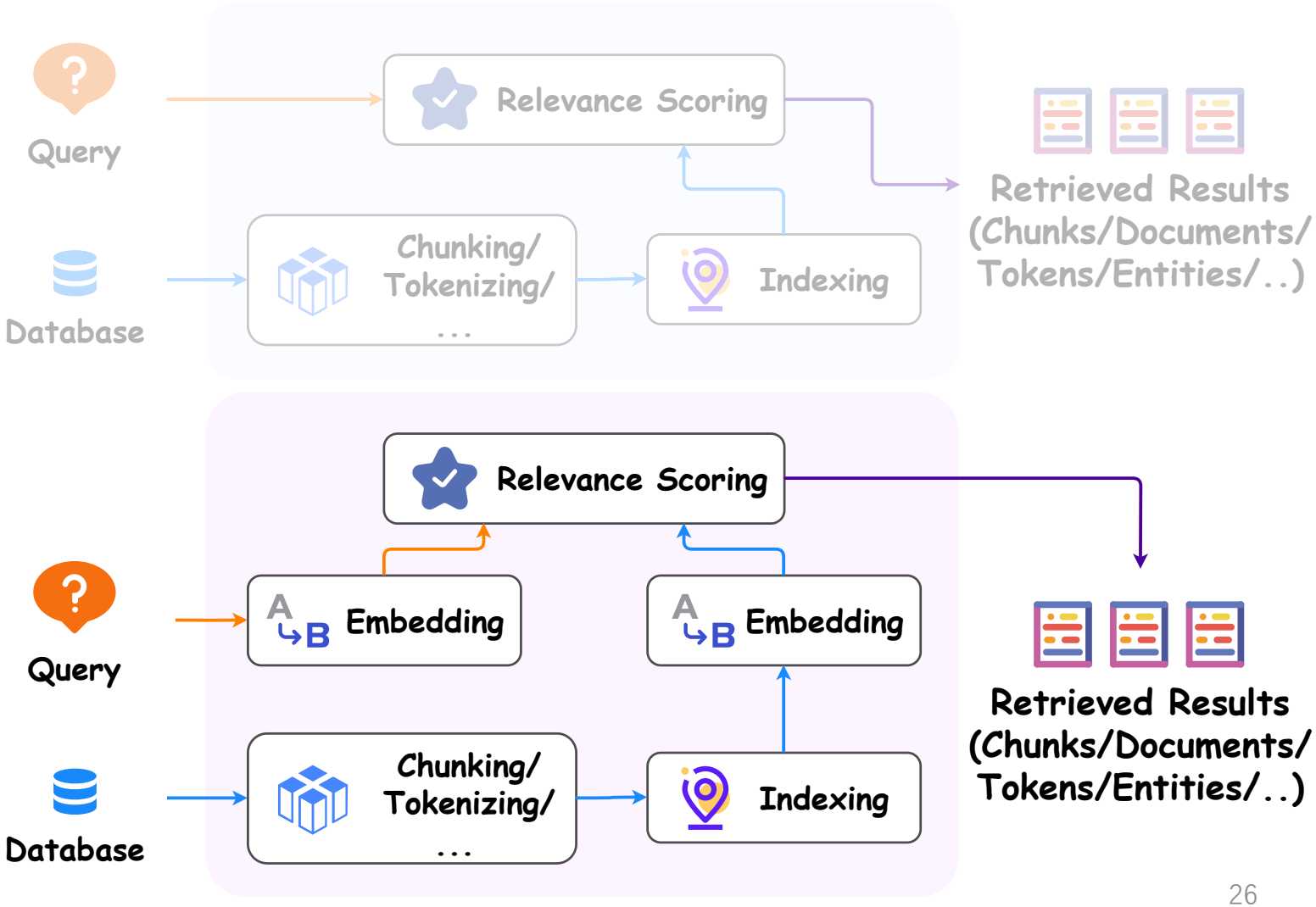
- Feasible to apply
- High efficiency
- Fine performance
- Example: TF-IDF, BM25



Dense v.s. Sparse Retrievers

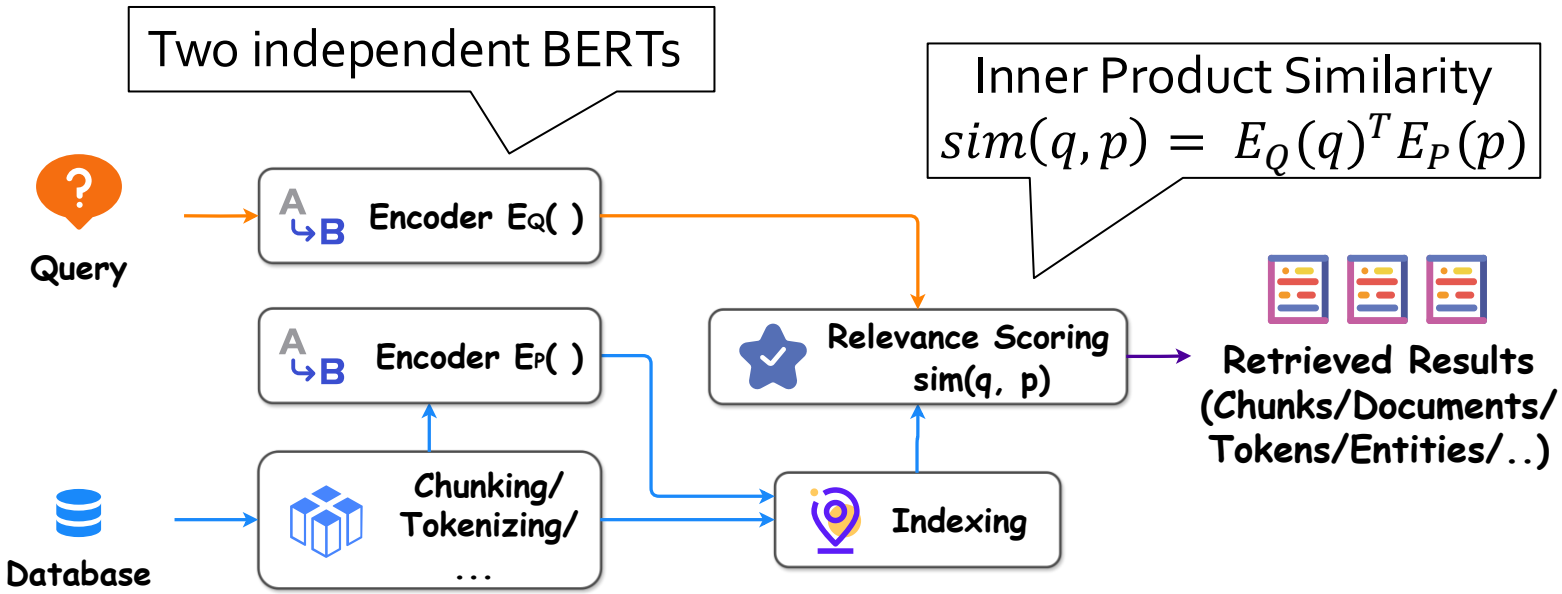
Dense Retrievers (DR)

- Allowing fine-tuning
- Better adaptation
- Customizable for more retrieval goals
- Example: DPR, Contriever, ColBERT



Task-Specific Retriever (Supervised)

- **Dense Passage Retriever (DPR):** Pretrained for Question Answering (QA)



Karpukhin et al. 2020. "Dense Passage Retrieval for Open-Domain Question Answering"

Task-Specific Retriever (Supervised)

- Dense Passage Retriever (DPR): **Pretrained for Question Answering (QA)**

  - Learning Objective

$$L(q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^-)$$
$$= -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^{n} e^{\text{sim}(q_i, p_{i,j}^-)}}$$

  - Training data: Question-Passage Sets

$$\mathcal{D} = \{\langle q_i, p_i^+, p_{i,1}^-, \cdots, p_{i,n}^- \rangle\}_{i=1}^{m}$$

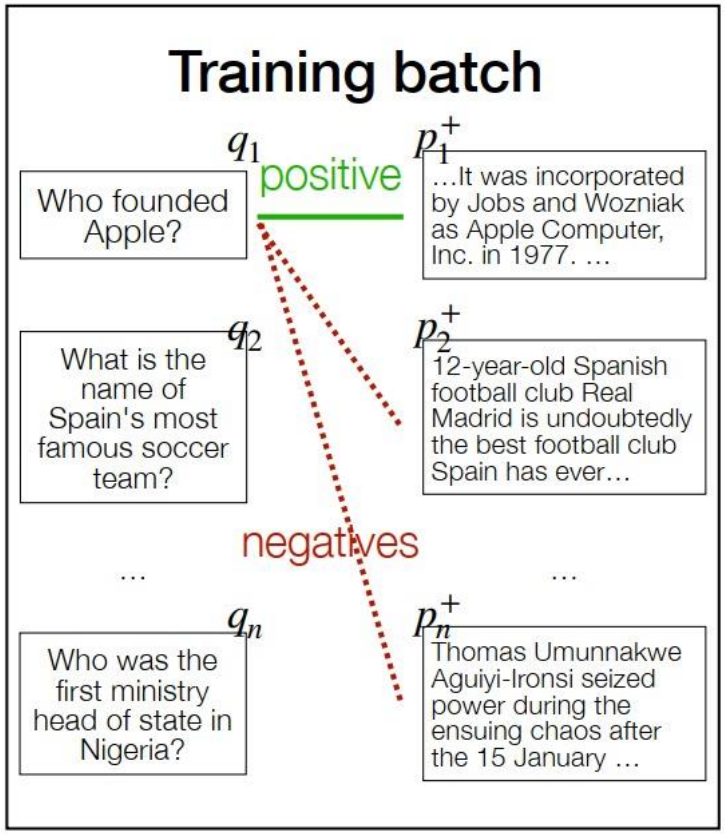
Question

Relevant passage

Irrelevant passages

Negative sample selection?

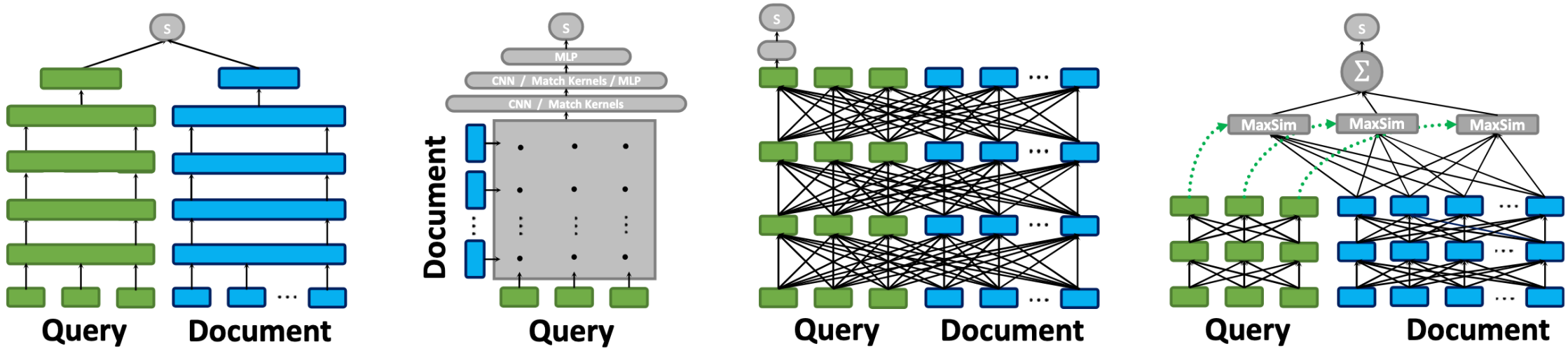
  - Training with in-batch negatives



Karpukhin et al. 2020. "Dense Passage Retrieval for Open-Domain Question Answering"

Task-Specific Retriever (Supervised)

❑ **ColBERT & ColBERTv2** : Late Interaction Retrieval



(a) Representation-based Similarity *(e.g., DSSM, SNRM)*

(b) Query-Document Interaction *(e.g., DRMM, KNRM, Conv-KNRM)*

(c) All-to-all Interaction *(e.g., BERT)*

(d) Late Interaction *(i.e., the proposed ColBERT)*

- Late Interaction

$$\mathrm{LI}(q,d) = \sum_{i \in [|1,N_q|]} \max_{j \in [|1,N_d|]} \langle \mathbf{E_q}^{(i)} | \mathbf{E_d}^{(j)} \rangle$$

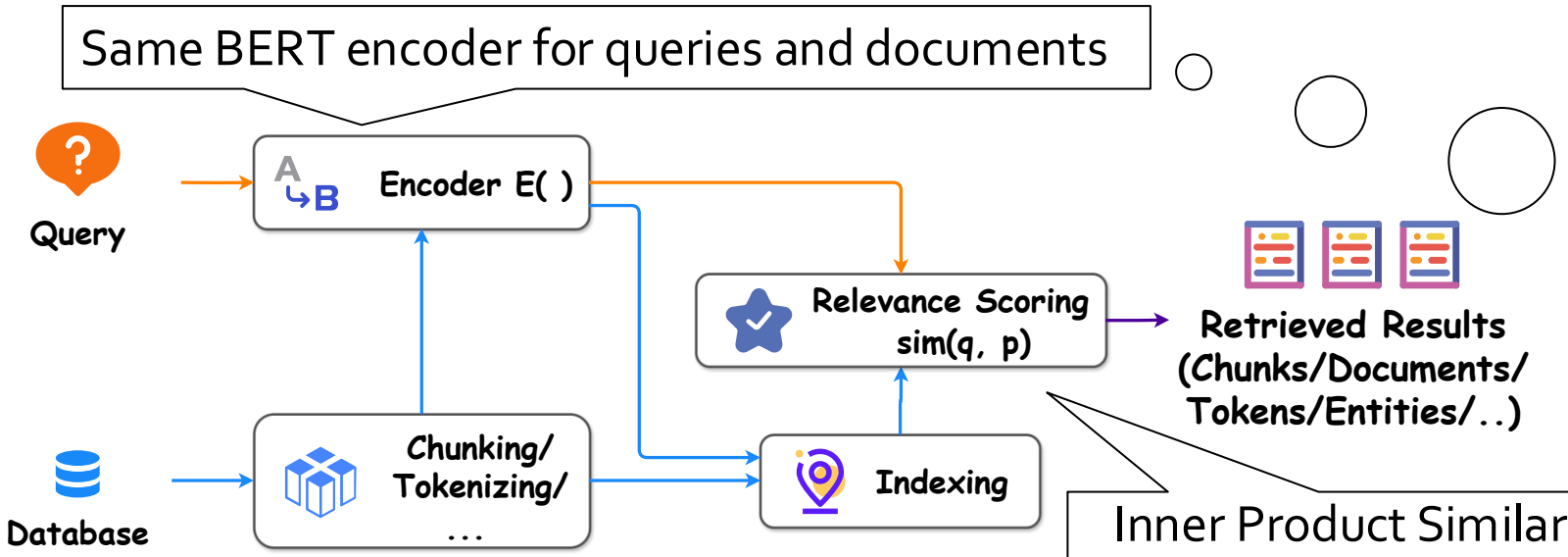
- Contrastive Loss

$$\mathcal{L} = -\frac{1}{b} \sum_{k=1}^{b} \log \left[\frac{\exp\left(s_k^+\right)}{\exp\left(s_k^+\right) + \exp\left(s_k^-\right)} \right]$$

Khattab et al. 2020. "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT". SIGIR 2020
Santhanam et al. 2022. "ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction". NAACL 2022

General-Purpose Retriever (Unsupervised)

- **Contriever:** Pre-trained with unsupervised learning
  - supervised pretrained methods do not transfer well to new applications with no training data



Might improve robustness in the context of zero-shot transfer or few-shot learning, while having no impact on other settings

Contrastive learning with **unaligned documents**

$$\mathcal{L}(q, k_+) = -\frac{\exp(s(q, k_+)/\tau)}{\exp(s(q, k_+)/\tau) + \sum_{i=1}^{K} \exp(s(q, k_i)/\tau)}$$

k_+: positive pairs from a single document

$(k_i)_{1,2,...,K}$: negative pairs with in-batch negative sampling and MoCo

Izacard et al. 2022. "Unsupervised Dense Information Retrieval with Contrastive Learning"

DPR & Contriever Performance on OpenQA Tasks

End-to-end QA (Exact Match) Accuracy

Training	Model	NQ	TriviaQA	WQ	TREC	SQuAD
Single	BM25+BERT (Lee et al., 2019)	26.5	47.1	17.7	21.3	33.2
Single	ORQA (Lee et al., 2019)	33.3	45.0	36.4	30.1	20.2
Single	HardEM (Min et al., 2019a)	28.1	50.9	-	-	-
Single	GraphRetriever (Min et al., 2019b)	34.5	56.0	36.4	-	-
Single	PathRetriever (Asai et al., 2020)	32.6	-	-	-	**56.5**
Single	$\text{REALM}_{\text{Wiki}}$ (Guu et al., 2020)	39.2	-	40.2	46.8	-
Single	$\text{REALM}_{\text{News}}$ (Guu et al., 2020)	40.4	-	40.7	42.9	-
Single	BM25	32.6	52.4	29.9	24.9	38.1
Single	DPR	**41.5**	56.8	34.6	25.9	29.8
Single	BM25+DPR	39.0	57.0	35.2	28.0	36.7
Multi	DPR	**41.5**	56.8	**42.4**	49.4	24.1
Multi	BM25+DPR	38.8	**57.9**	41.1	**50.6**	35.8

Both widely applied in RAG and RA-LFMs

DPR in

RAG, FiD, RETRO, EPR, UDR, …

Contriever in

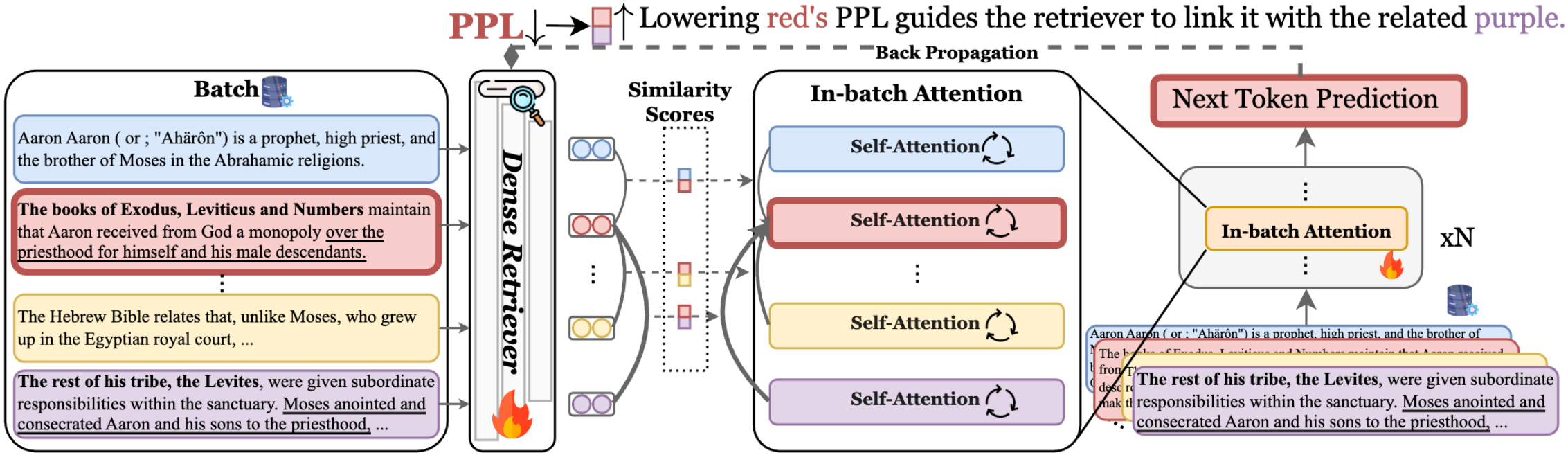
Self-RAG, Atlas, RAVEN, …

Both better than the sparse retriever!

	NaturalQuestions			TriviaQA		
	R@5	R@20	R@100	R@5	R@20	R@100
Inverse Cloze Task (Sachan et al., 2021)	32.3	50.9	66.8	40.2	57.5	73.6
Masked salient spans (Sachan et al., 2021)	41.7	59.8	74.9	53.3	68.2	79.4
BM25 (Ma et al., 2021)	-	62.9	78.3	-	**76.4**	**83.2**
Contriever	**47.8**	**67.8**	**82.1**	**59.4**	74.2	**83.2**
supervised model: DPR (Karpukhin et al., 2020)	-	78.4	85.4	-	79.4	85.0

General-Purpose Retriever (Self-Supervised)

❑ **Revela :** From contrastive self-supervision to LM-based retriever learning.

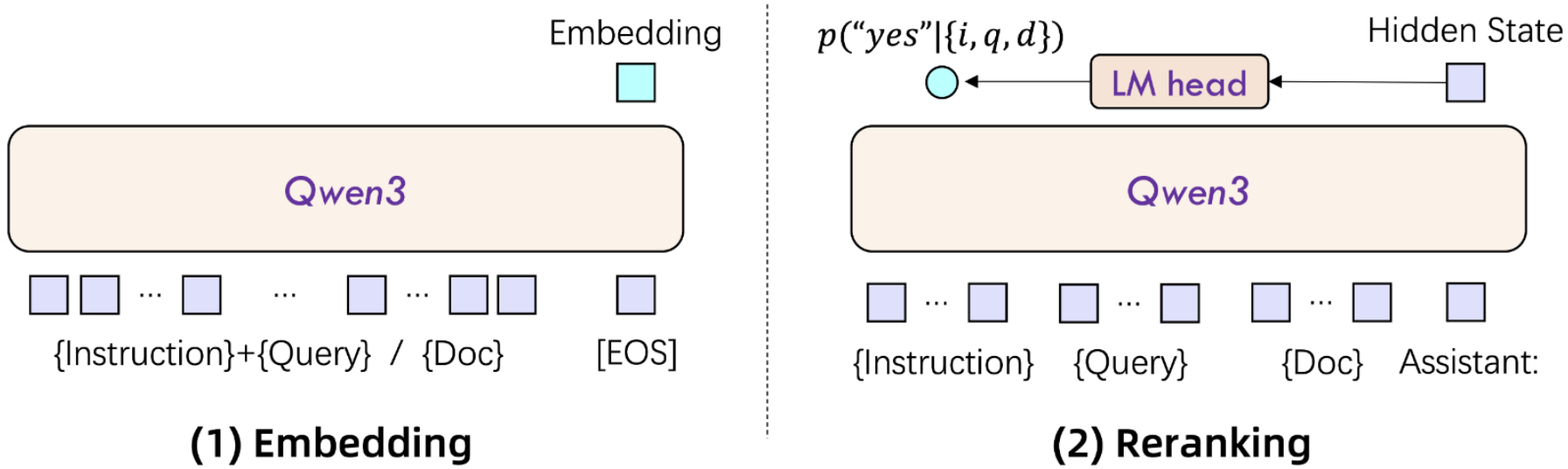


- Learns dense retrieval from raw text
- No query-document labels
- Retriever scores guide in-batch attention
- Optimized by next-token prediction

Cai et al. 2026. "Revela: Dense Retriever Learning via Language Modeling" ICLR 2026.

General-Purpose Retriever (LFM-based)

- **Qwen3 Embedding:** Text Embedding based on Foundation Models



- Training Objective

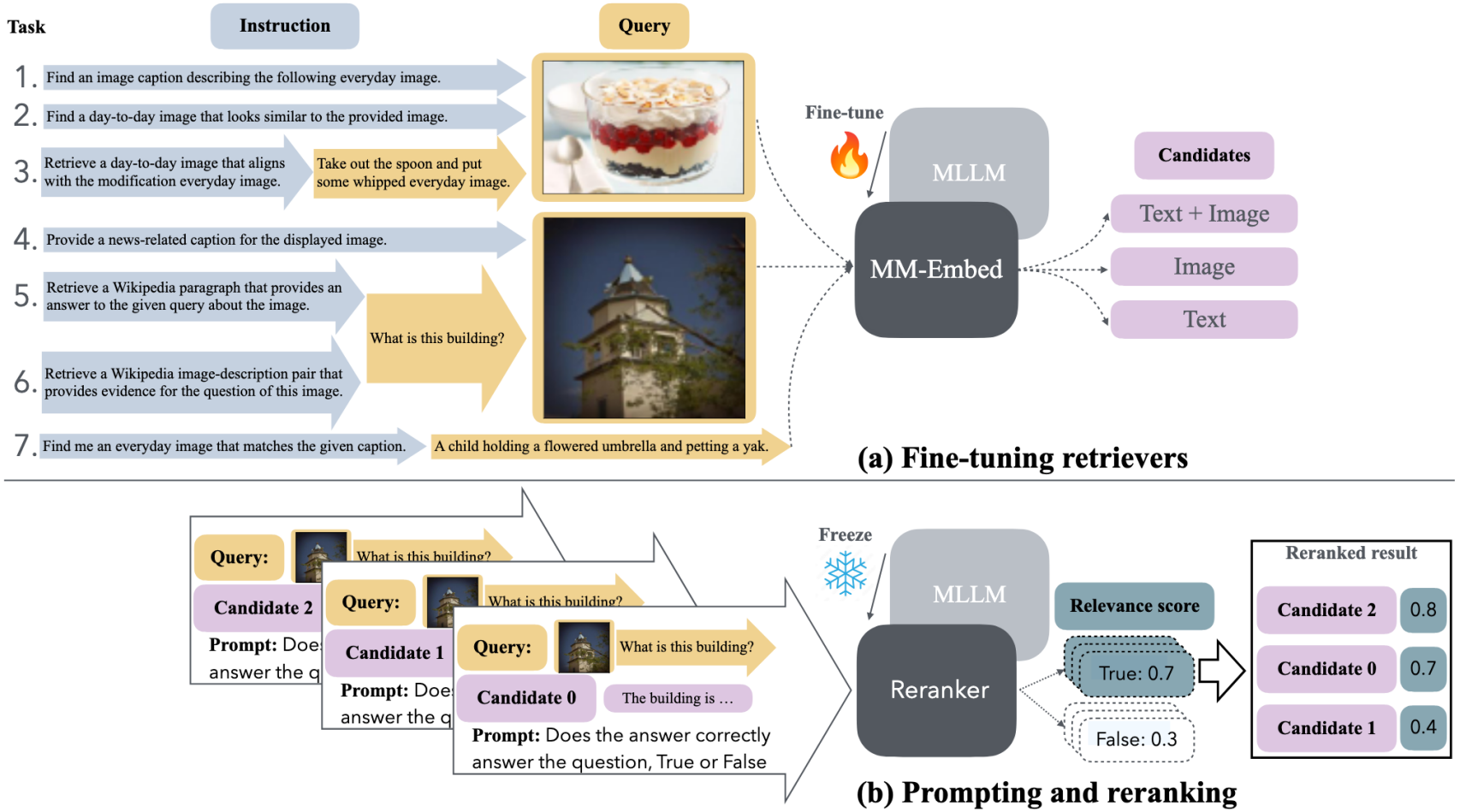
$$L_{\text{embedding}} = -\frac{1}{N}\sum_{i}^{N}\log\frac{e^{(s(q_i,d_i^+)/\tau)}}{Z_i}$$

$$L_{\text{reranking}} = -\log p(l|\mathcal{P}(q,d))$$

Zhang et al. 2025. “Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models.”

General-Purpose Retriever (LFM-based)

- **MM-EMBED:** Universal Multimodal Retrieval with Instruction-aware Capability



Task	
1. $q^{\text{txt}} \rightarrow c^{\text{img}}$	
2. $q^{\text{txt}} \rightarrow c^{\text{txt}}$	
3. $q^{\text{txt}} \rightarrow (c^{\text{img}}, c^{\text{txt}})$	
4. $q^{\text{img}} \rightarrow c^{\text{txt}}$	
5. $q^{\text{img}} \rightarrow c^{\text{img}}$	
6. $(q^{\text{img}}, q^{\text{txt}}) \rightarrow c^{\text{txt}}$	
7. $(q^{\text{img}}, q^{\text{txt}}) \rightarrow c^{\text{img}}$	
8. $(q^{\text{img}}, q^{\text{txt}}) \rightarrow (c^{\text{img}}, c^{\text{txt}})$	

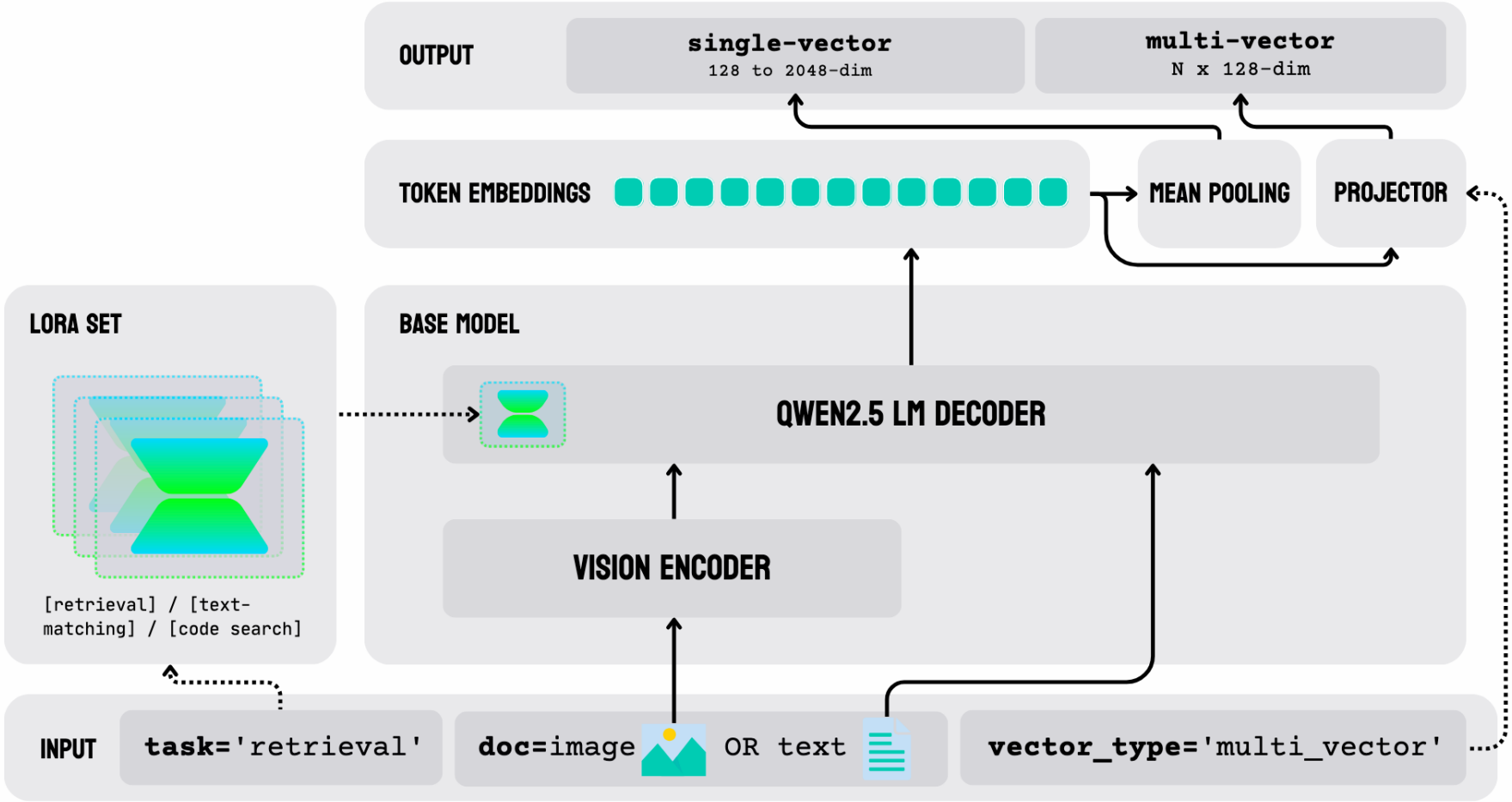
Lin et al. 2025. "MM-Embed: Universal Multimodal Retrieval with Multimodal LLMs." ICLR 2025.

General-Purpose Retriever (LFM-based)

❑ **jina-embeddings-v4:** Universal Multi-task Multi-modal Embedding Model

Task-specific LoRA adapters

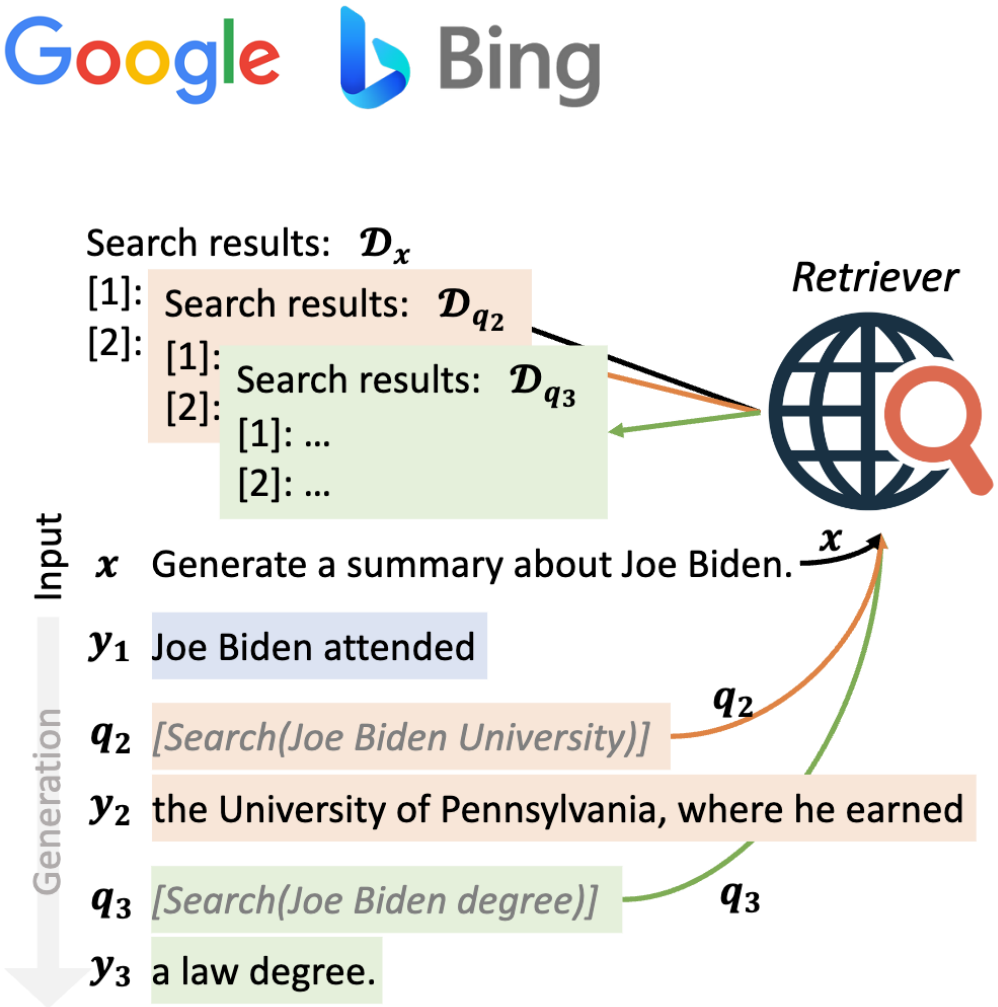
- ❖ Retrieval
- ❖ Text Matching
- ❖ Code Search
- ❖ …



Günther et al. 2025. "jina-embeddings-v4: Universal Embeddings for Multimodal Multilingual Retrieval." EMNLP 2025.

Search Engine as Retrievers

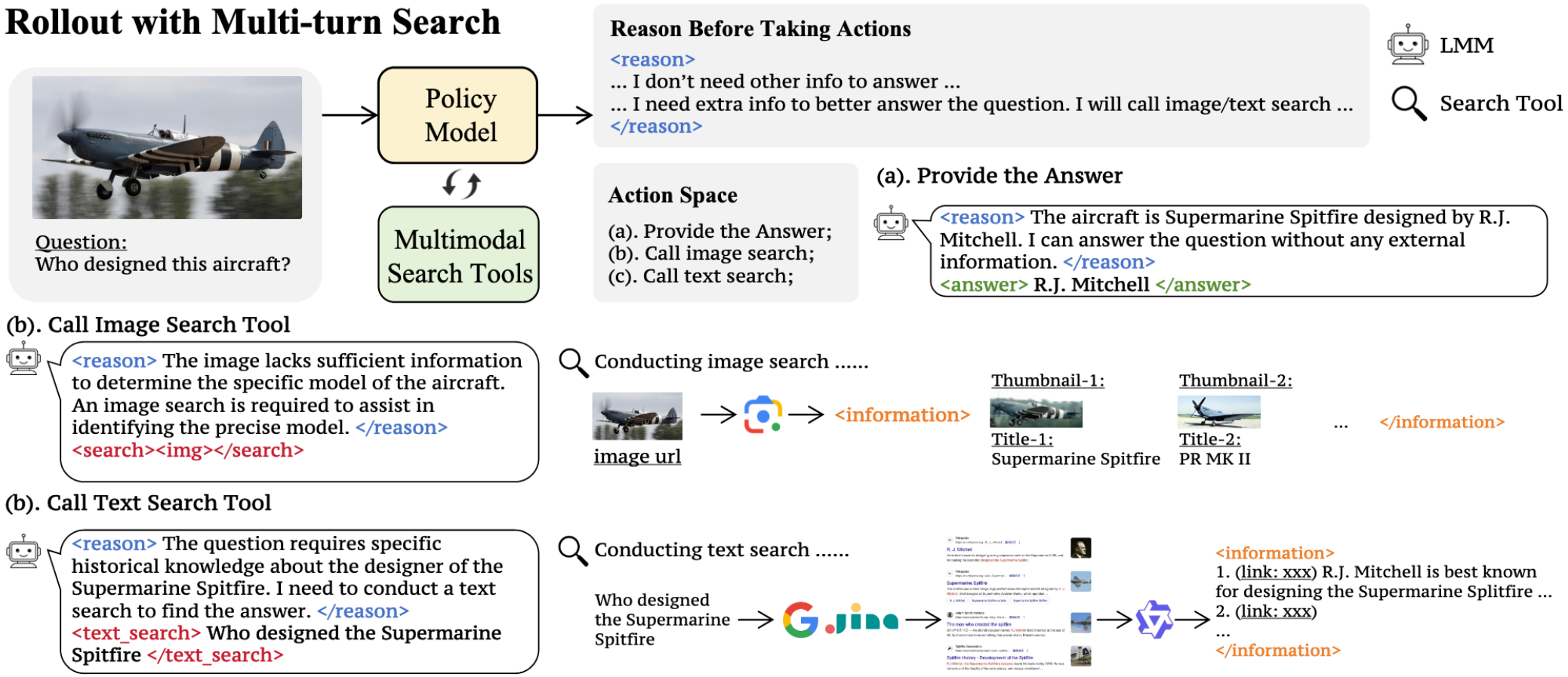
- **Traditional retrieval methods**
  - May be difficult to update to real-time web documents
  - May be a limit to the number of documents storable in the pre-defined database
  - Will not take advantage of the high quality ranking that has been finely tuned in Internet Search engines over decades of use



Komeili et al., 2021, Internet-Augmented Dialogue Generation
Jiang et al. 2023, Active Retrieval Augmented Generation

Search Engine as Retrievers

- **MMSearchR1:** Search engines as reasoning-time retrievers.



Wu et al., 2026, "Multimodal-Search-R1: Incentivizing LMMs to Search." ACL 2026

PART 2: Architecture of RA-LFMs and Main Modules

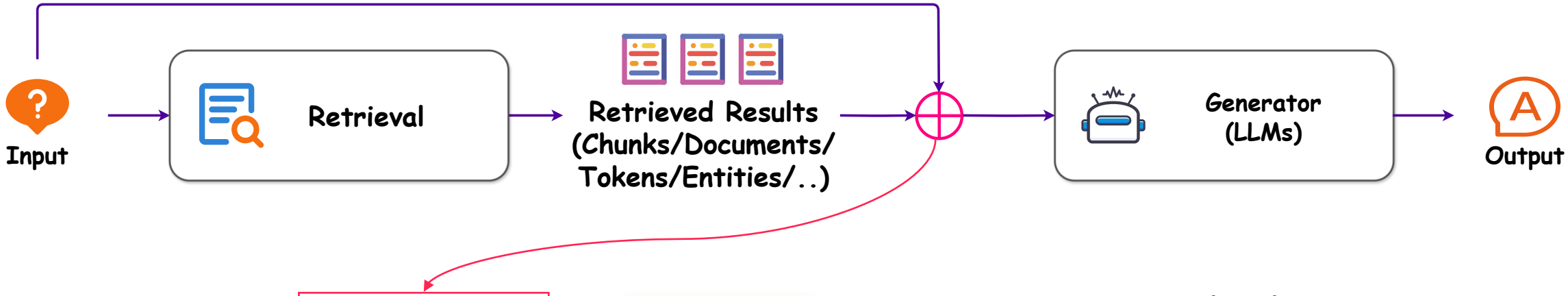

Slides

Website of this tutorial

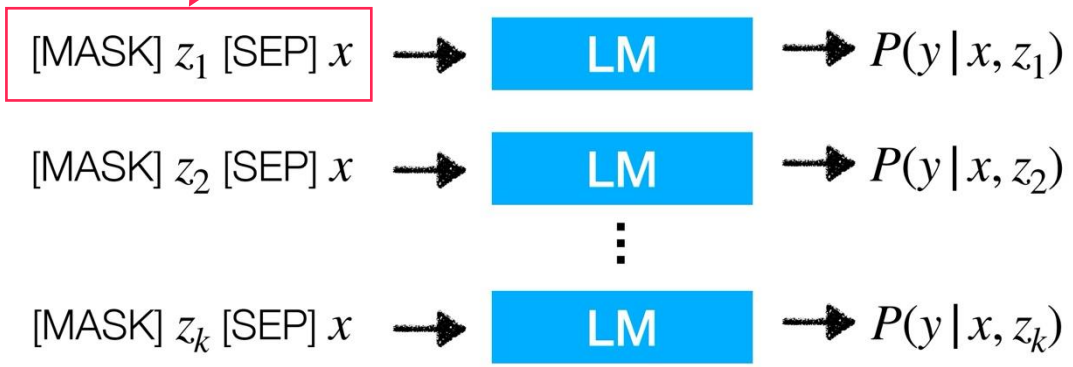
- **RA-LFM architecture overview**
- **Retriever in RA-LFMs**
- **Retrieval results integration**
- **Pre/Post-retrieval techniques**
- **Special RA-LFM paradigms**

Retrieved Results Integration: Input-layer Integration

❑ REALM



Integrating the retrieved passage $\boldsymbol{z}$ and the original input $\boldsymbol{x}$



Weighted aggregating the prediction results based on all retrieved passages

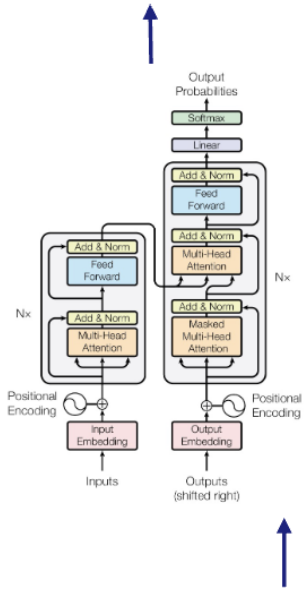
$$\sum_{z \in \mathcal{D}} P(z \mid x) P(y \mid x, z)$$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training." ICML 2020

Retrieval-Augmented Generator

Typical encoder: $p(y|x)$

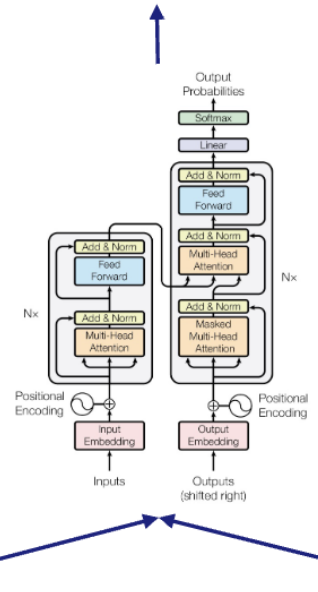
y= pounds



x: we paid 20 __ at the Buckingham Palace gift shop

Knowledge-augmented encoder:$p(y|x,z)$

y= pounds

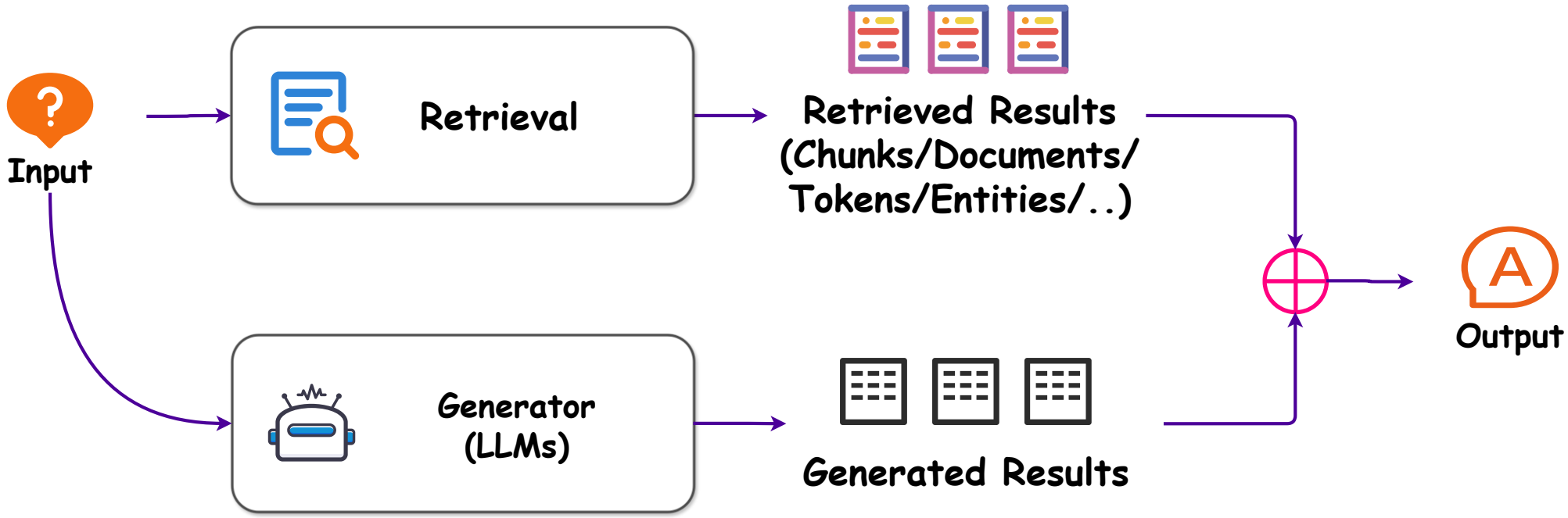


x: we paid 20 __ at the Buckingham Palace gift shop

z: Buckingham Palace is home to the British monarchy

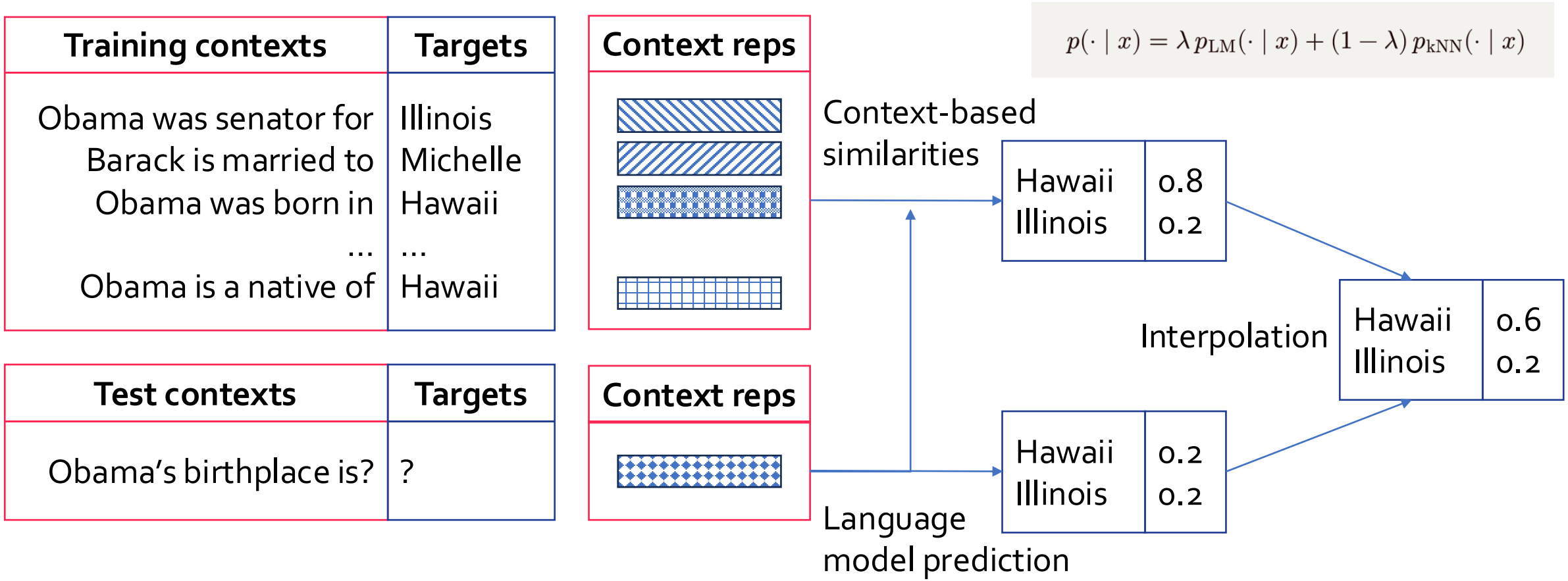
explicit knowledge

Retrieved Results Integration: Output-layer integration



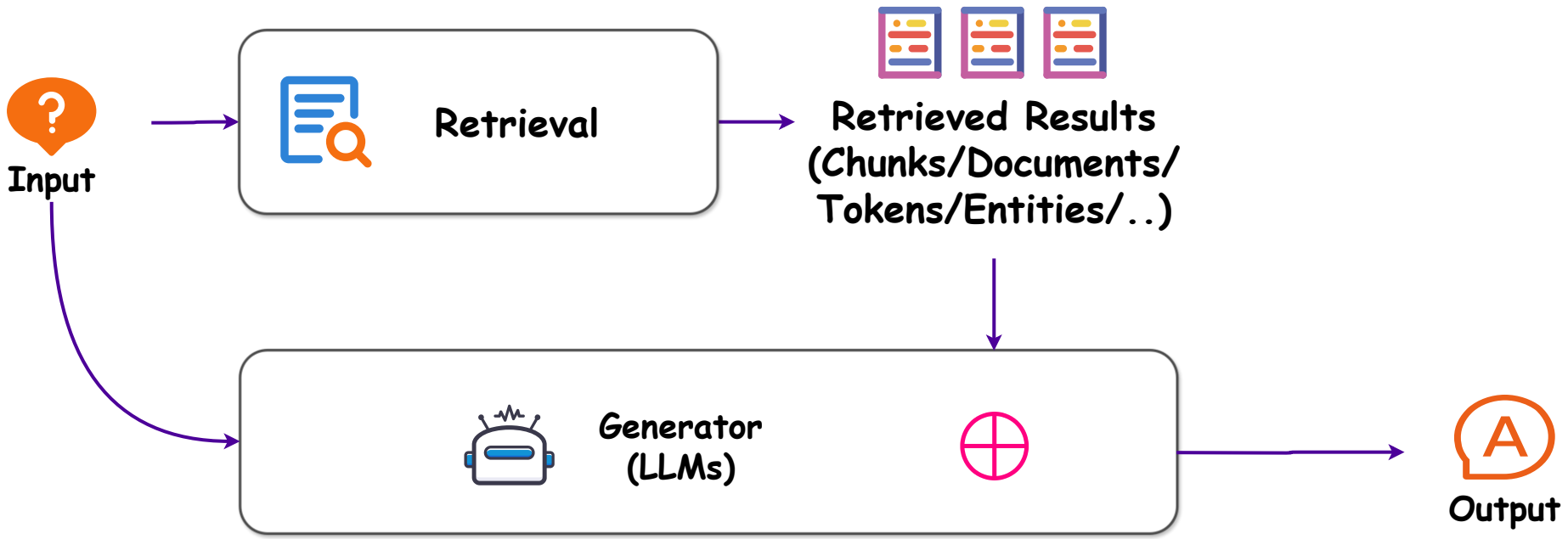
RA-LFM Architecture: Output-layer Integration

- **kNN-LM:** Combining retrieved probabilities and predicted ones in generation



Khandelwal el al. 2020. "Generalization through Memorization: Nearest Neighbor Language Models." ICLR 2020

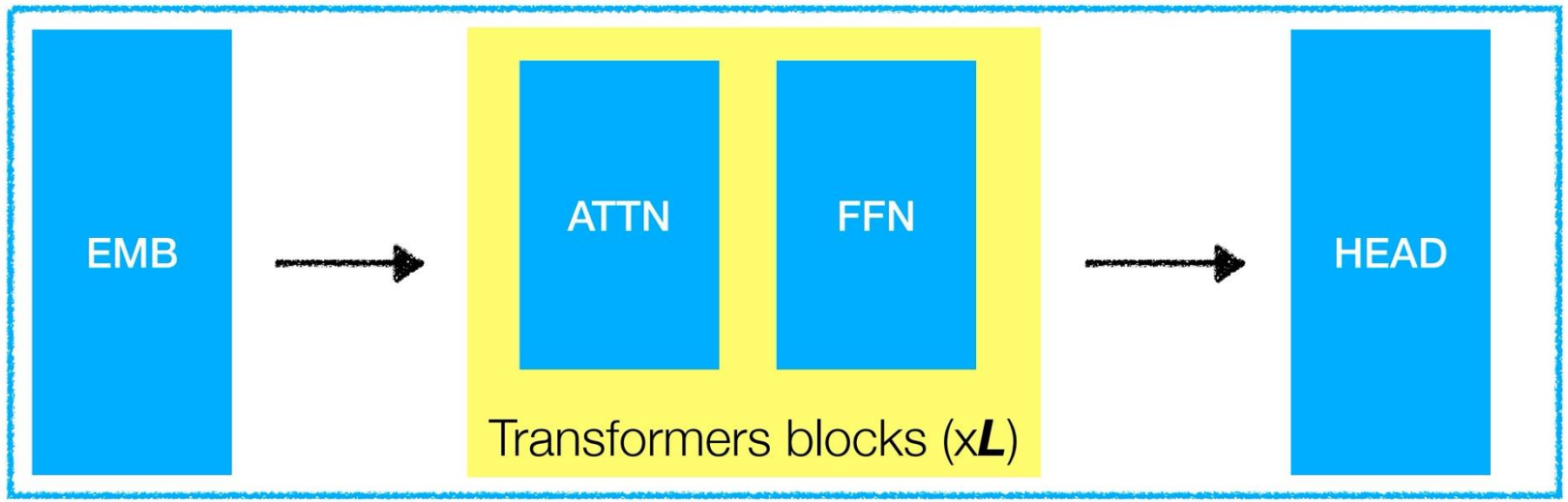
Retrieved Results Integration: Intermediate-layer Integration



Borgeaud et al. 2022. “Improving Language Models by Retrieving from Trillions of Tokens.” ICML 2022

Retrieved Results Integration: Intermediate-layer Integration

Regular Decoder

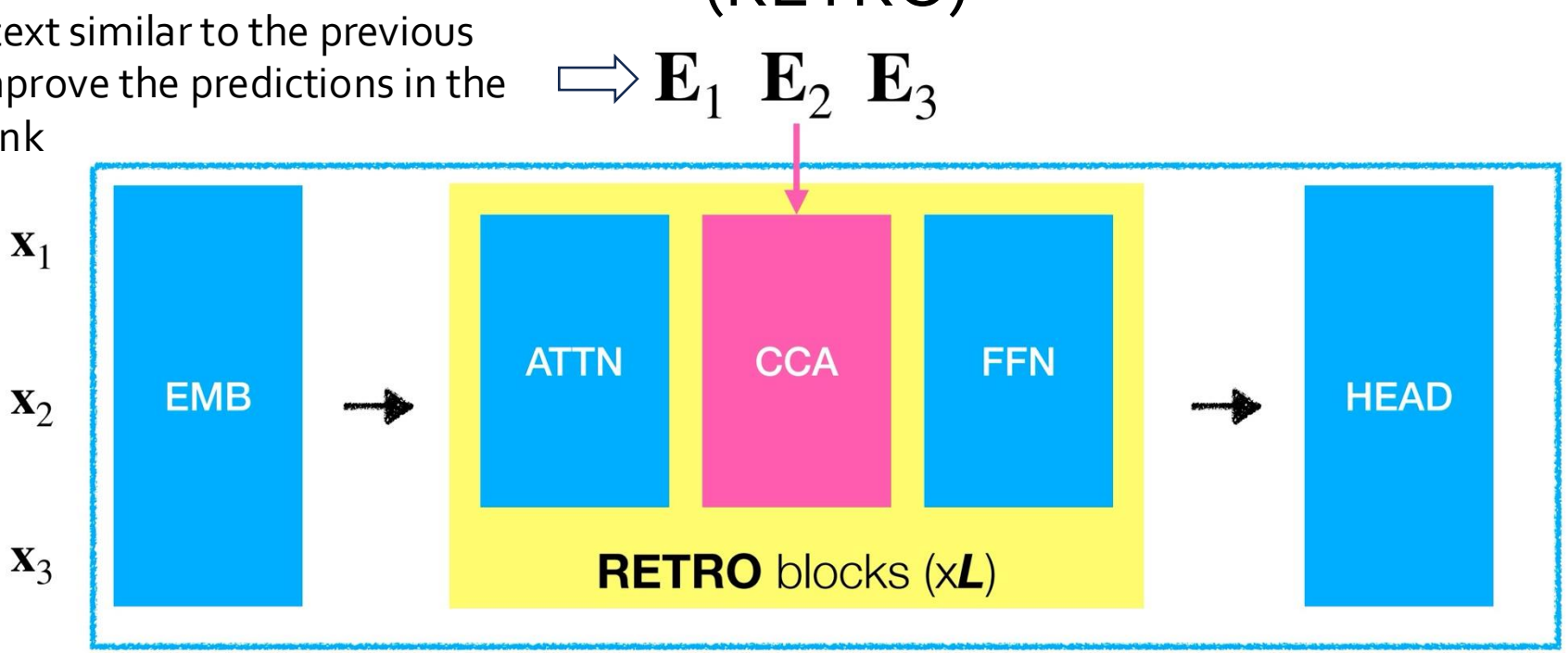


Borgeaud et al. 2022. "Improving Language Models by Retrieving from Trillions of Tokens." ICML 2022

Retrieved Results Integration: Intermediate-layer Integration

Decoder to incorporate retrieved results (RETRO)

Retrieving text similar to the previous chunk to improve the predictions in the current chunk

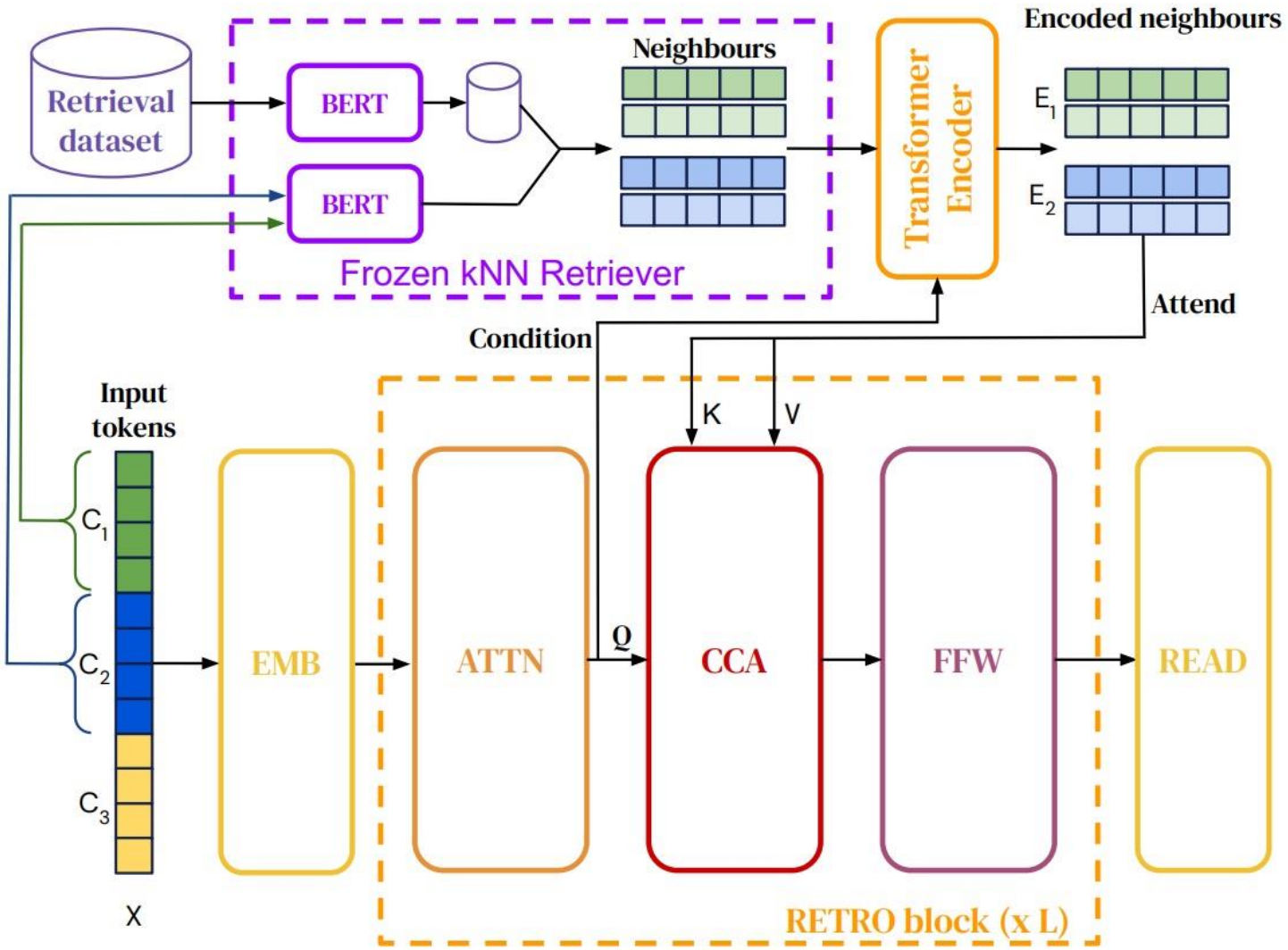


Spliting the input sequence into chunks

Chunked Cross Attention (CCA)

Borgeaud et al. 2022. "Improving Language Models by Retrieving from Trillions of Tokens." ICML 2022

Retrieved Results Integration: Intermediate-layer Integration



Borgeaud et al. 2022. "Improving Language Models by Retrieving from Trillions of Tokens." ICML 2022

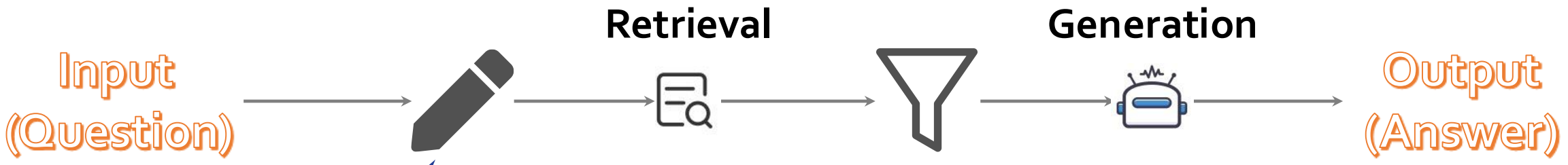
PART 2: Architecture of RA-LFMs and Main Modules


Slides

Website of this tutorial

- **RA-LFM architecture overview**
- **Retriever in RA-LFMs**
- **Retrieval results integration**
- **Pre/Post-retrieval techniques**
- **Special RA-LFM paradigms**

Pre/Post-Retrieval Techniques



Pre-retrieval process: to improve the adaptation and effectiveness of the query

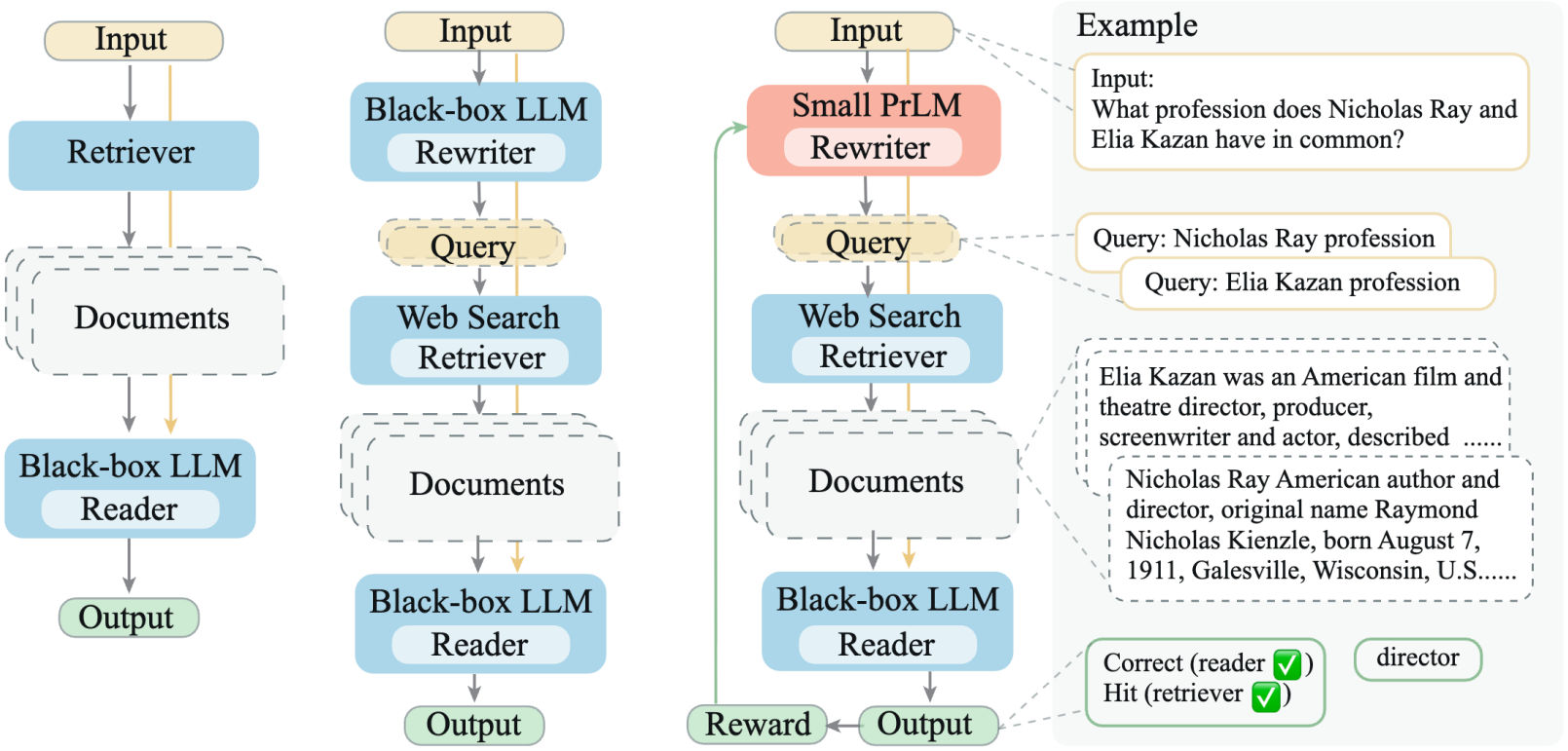
- Query rewriting
- Query decomposition
- Query expansion

Post-retrieval process: to select better results, merge multiple ones, etc

- Reranking
- Pruning
- Verification/Correction
- Dynamic Selection

Pre-Retrieval Techniques

- **Query Rewriting**: to improve the adaptation of the query



(a) Retrieve-then-read (b)Rewrite-retrieve-read (c) Trainable rewrite-retrieve-read

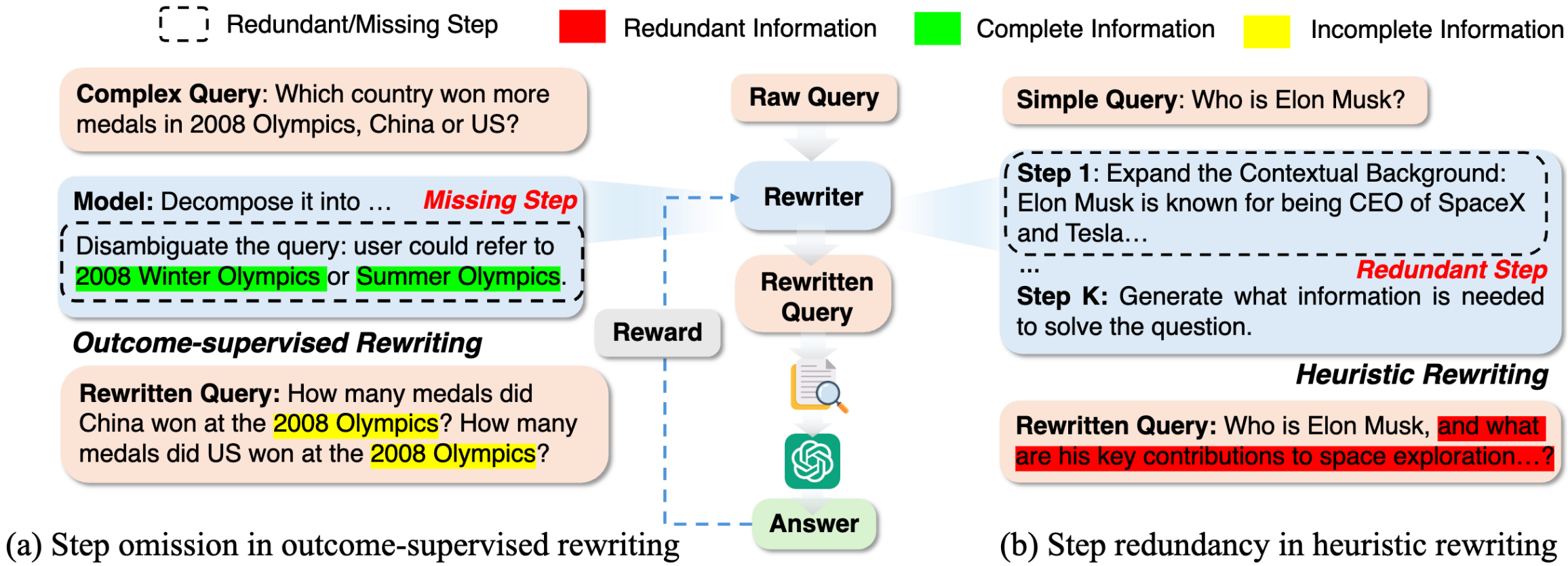
Model	EM	F_1
HotpotQA		
Direct	32.36	43.05
Retrieve-then-read	30.47	41.34
LLM rewriter	32.80	43.85
Trainable rewriter	34.38	45.97
AmbigNQ		
Direct	42.10	53.05
Retrieve-then-read	45.80	58.50
LLM rewriter	46.40	58.74
Trainable rewriter	47.80	60.71
PopQA		
Direct	41.94	44.61
Retrieve-then-read	43.20	47.53
LLM rewriter	46.00	49.74
Trainable rewriter	45.72	49.51

Works on different QA settings

Wang et al. 2023. "Query rewriting for retrieval-augmented large language models." EMNLP 2023

Pre-Retrieval Techniques

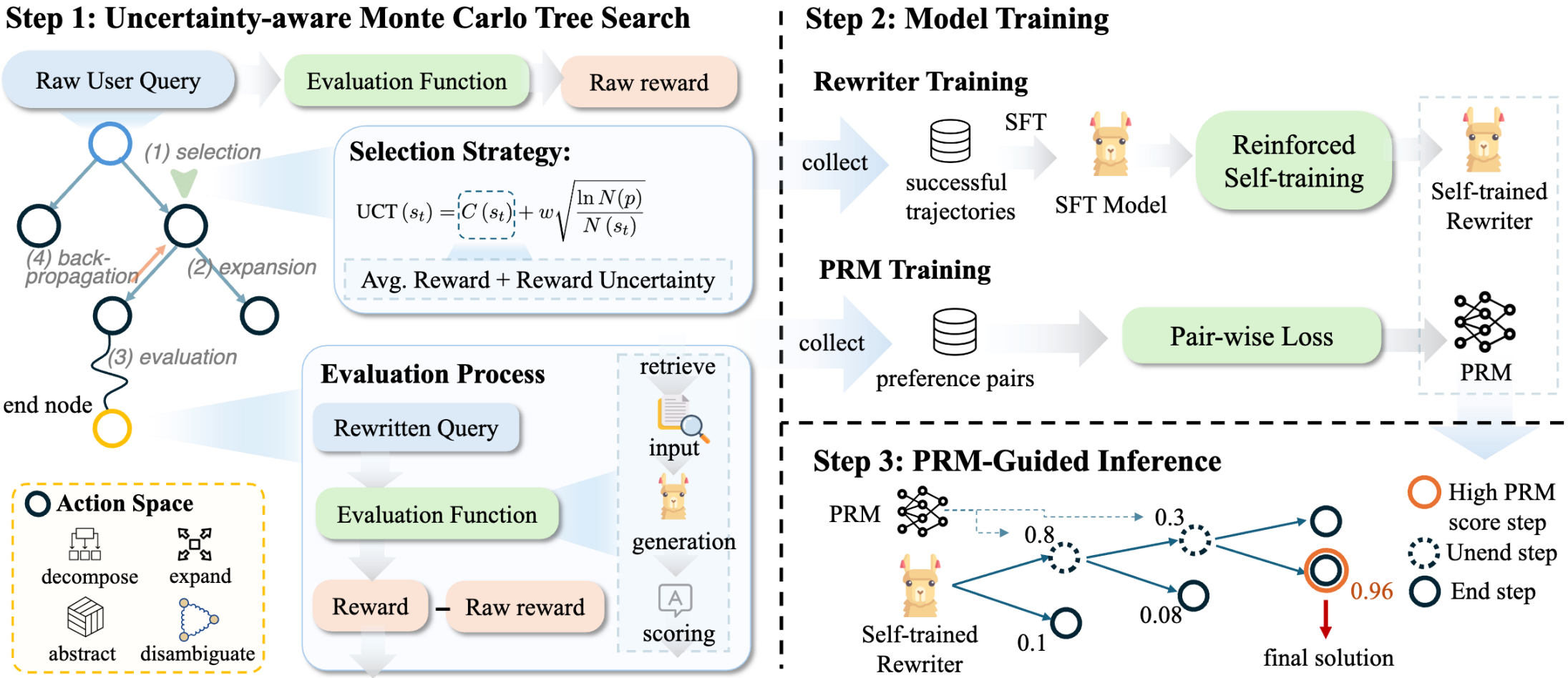
- **Adaptive Query Rewriting** : Step-level Process Supervision



Ye et al. 2025. "Q-PRM: Adaptive Query Rewriting for Retrieval-Augmented Generation via Step-level Process Supervision". EMNLP Findings 2025

Pre-Retrieval Techniques

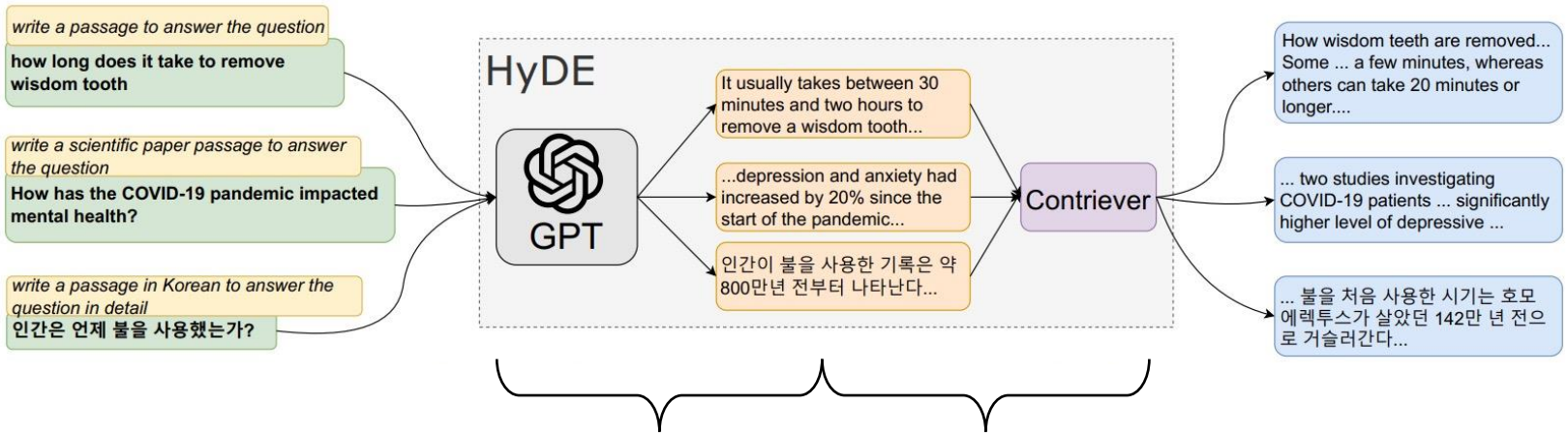
- **Adaptive Query Rewriting** : Step-level Process Supervision

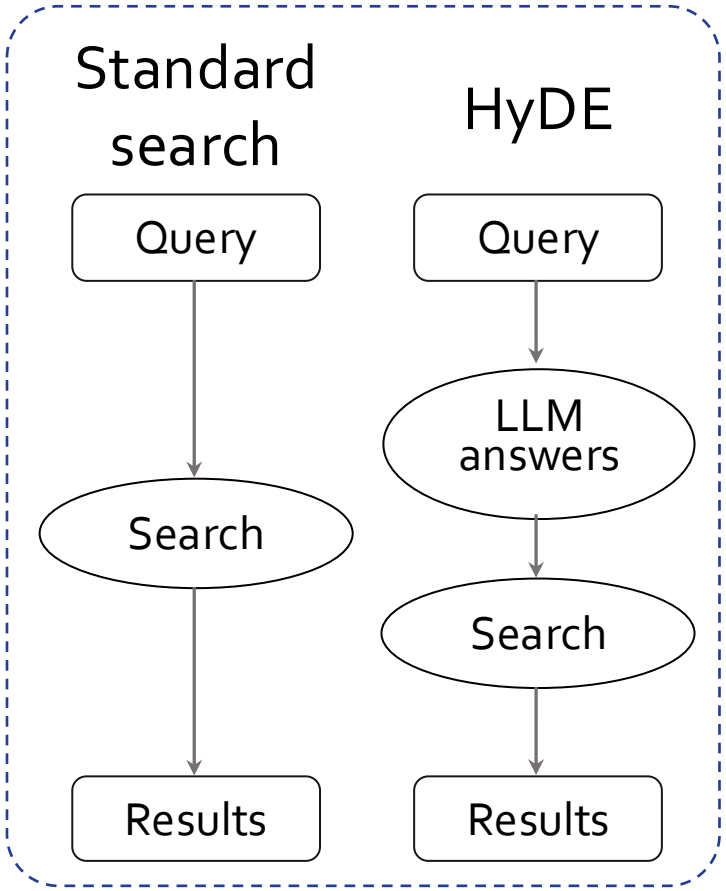


Ye et al. 2025. "Q-PRM: Adaptive Query Rewriting for Retrieval-Augmented Generation via Step-level Process Supervision". EMNLP Findings 2025

Pre-Retrieval Techniques

- **Query Expansion**: Hypothetical Document Embeddings (HyDE)

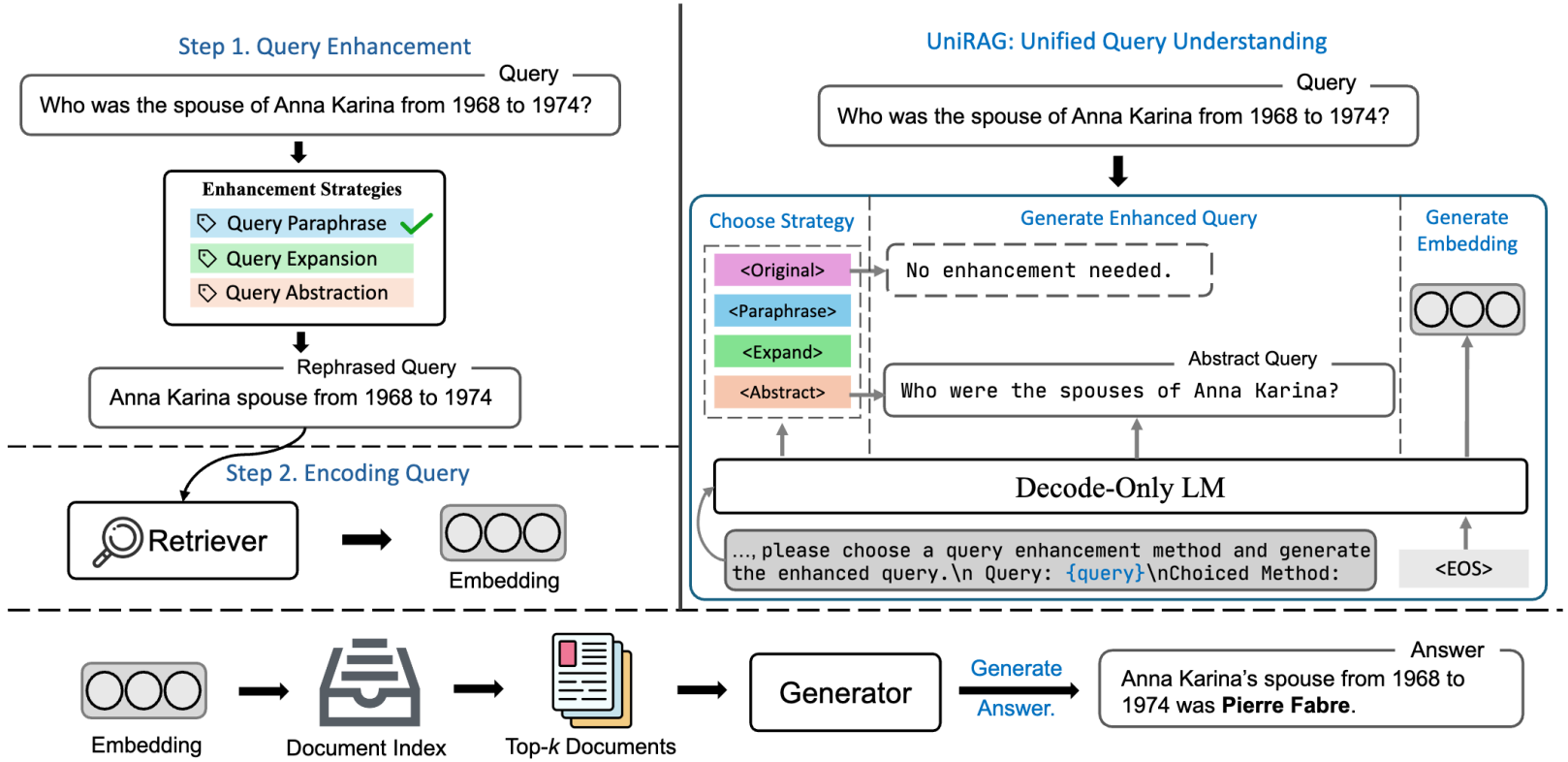


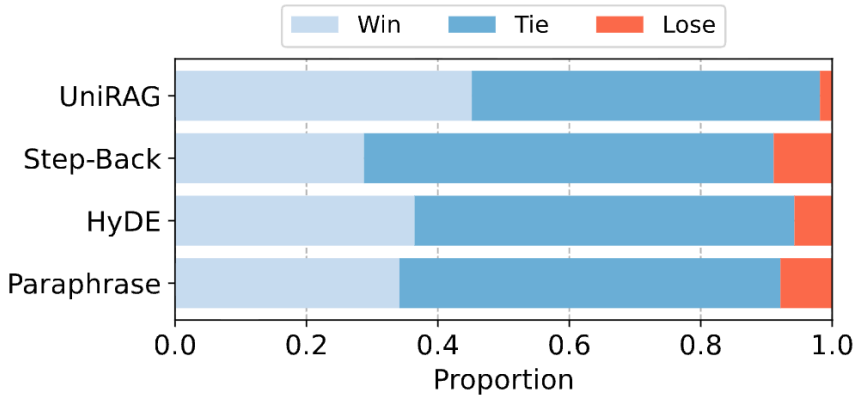


Gao et al. 2023. "Precise zero-shot dense retrieval without relevance labels." ACL 2023

Pre-Retrieval Techniques

❑ Query Expansion





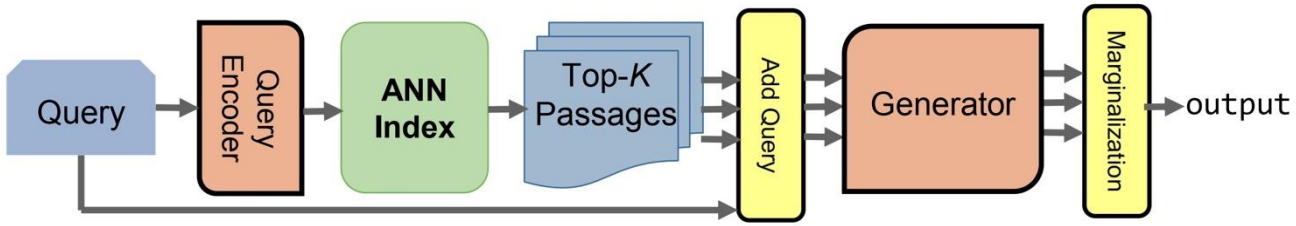
Comparison of win rates of different augmentation methods.

Li et al. 2025. "UniRAG: Unified Query Understanding Method for Retrieval Augmented Generation", ACL 2025.

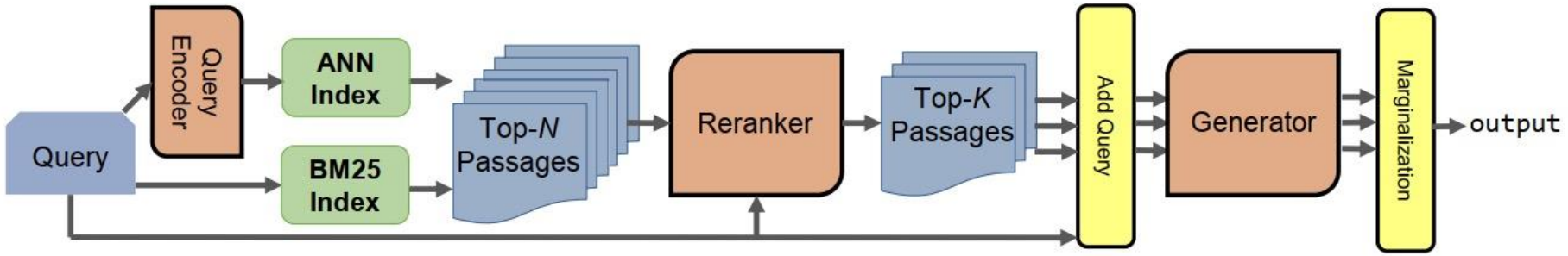
Post-Retrieval Techniques

❑ Retrieved Result Rerank (Re2G)

- ❖ Results from initial retrieval can be greatly improved through the use of a reranker
- ❖ Reranker allows merging retrieval results from sources with incomparable scores, e.g., BM25 and neural initial retrieval



Regular RAG pipeline

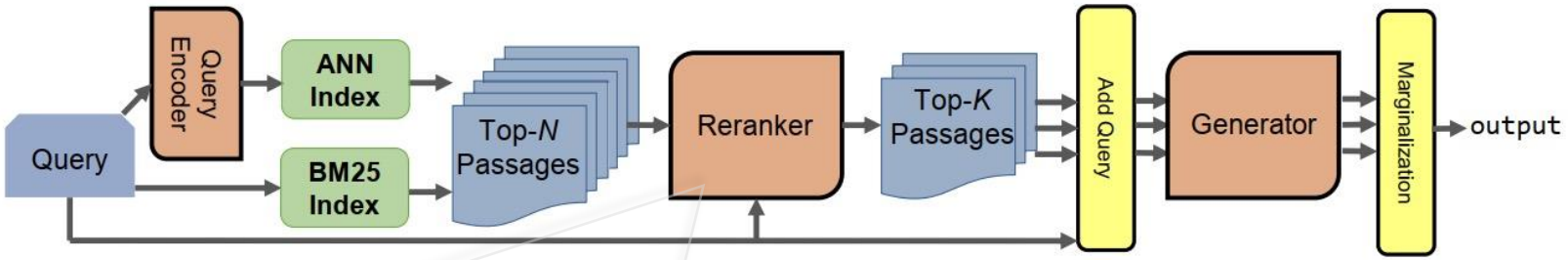


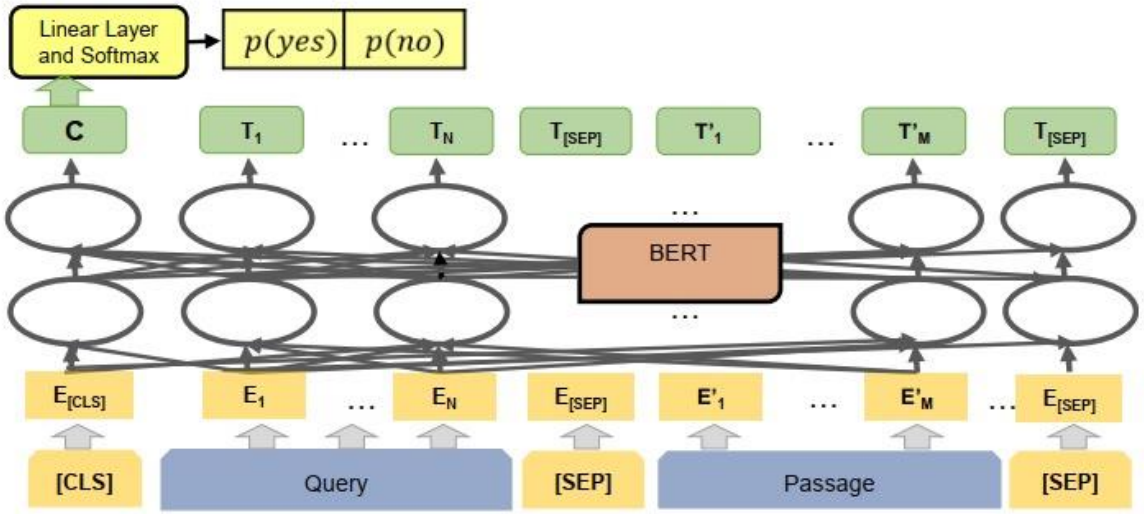
RAG with Reranker

Glass et al. 2022. "Re2G: Retrieve, Rerank, Generate." NAACL 2022

Retrieved Result Rerank Model

- **Reranker: interaction model based on the sequence-pair classification**





Nogueira and Cho, 2019, Passage Re-ranking with BERT

Retrieved Result Rerank (Re2G) Performance

	T-REx		NQ		TriviaQA		FEVER		WoW	
	R-Prec	**R@5**	**R-Prec**	**R@5**	**R-Prec**	**R@5**	**R-Prec**	**R@5**	**R-Prec**	**R@5**
BM25	46.88	69.59	24.99	42.57	26.48	45.57	42.73	70.48	27.44	45.74
DPR Stage 1	49.02	63.34	56.64	64.38	60.12	64.04	75.49	84.66	34.74	60.22
KGI_0 DPR	65.02	75.52	64.65	69.60	60.55	63.65	80.34	86.53	**48.04**	**71.02**
Re^2G DPR	**67.16**	**76.42**	**65.88**	**70.90**	**62.33**	**65.72**	**84.13**	**87.90**	47.09	69.88
KGI_0 DPR+BM25	60.48	80.06	36.91	66.94	40.81	64.79	65.95	90.34	35.63	68.47
Reranker Stage 1	81.22	87.00	70.78	73.05	**71.80**	**71.98**	87.71	92.43	55.50	**74.98**
Re^2G Reranker	**81.24**	**88.58**	**70.92**	**74.79**	60.37	70.61	**90.06**	**92.91**	**57.89**	74.62

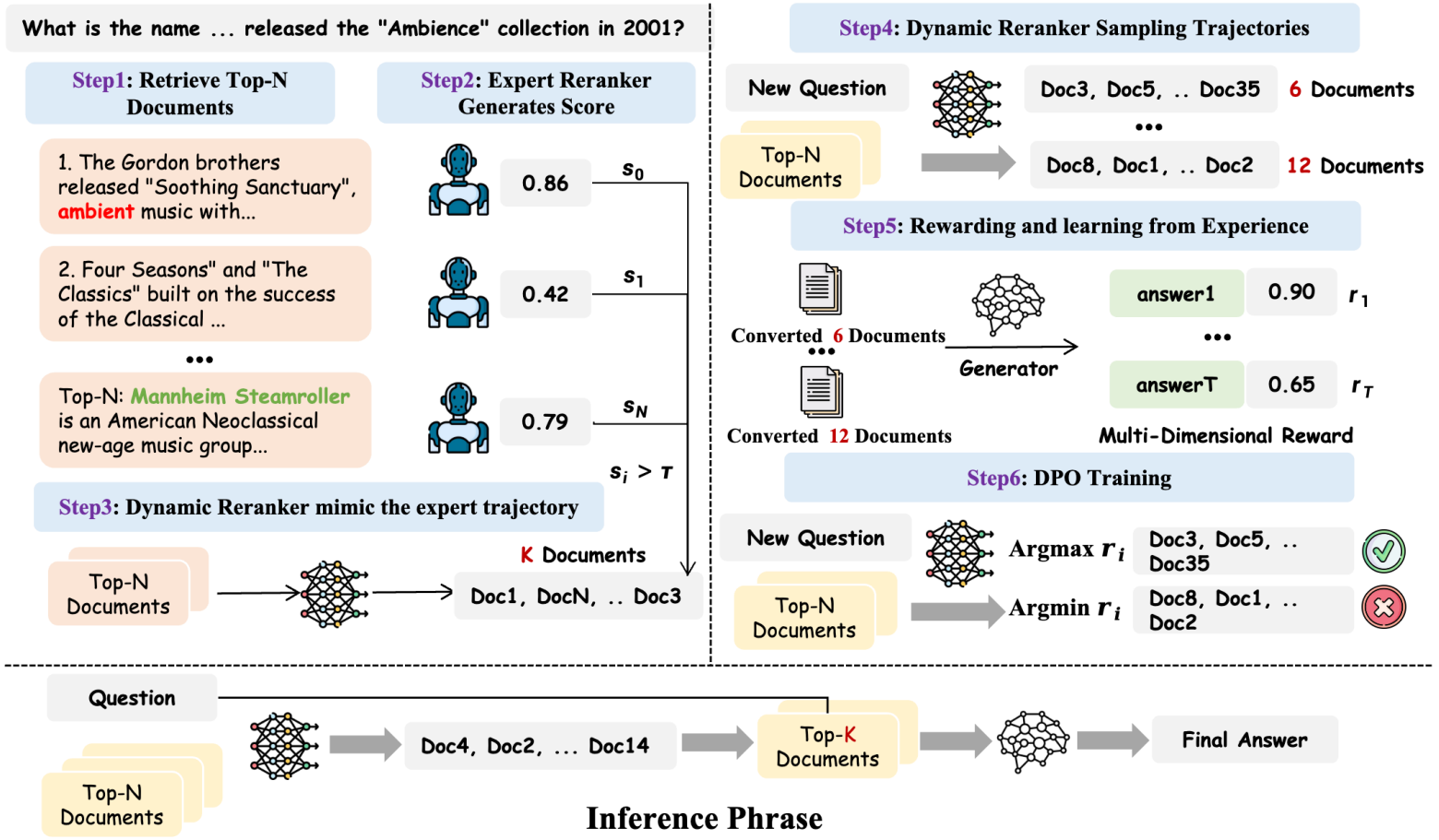
Significantly outperforms pipelines without the *Rerank* stage

Glass et al. 2022. "Re2G: Retrieve, Rerank, Generate"

Retrieved Result Rerank Model

❑ **DynamicRAG:** Dynamic Reranking in Retrieval-Augmented Generation

- ❖ **Question**: How many retrieved documents should be passed to the generator?
- ❖ **Problem**: Fixed top-k retrieval is query-agnostic.
- ❖ **Key idea**: Dynamically adjust document order and document number using generation feedback.

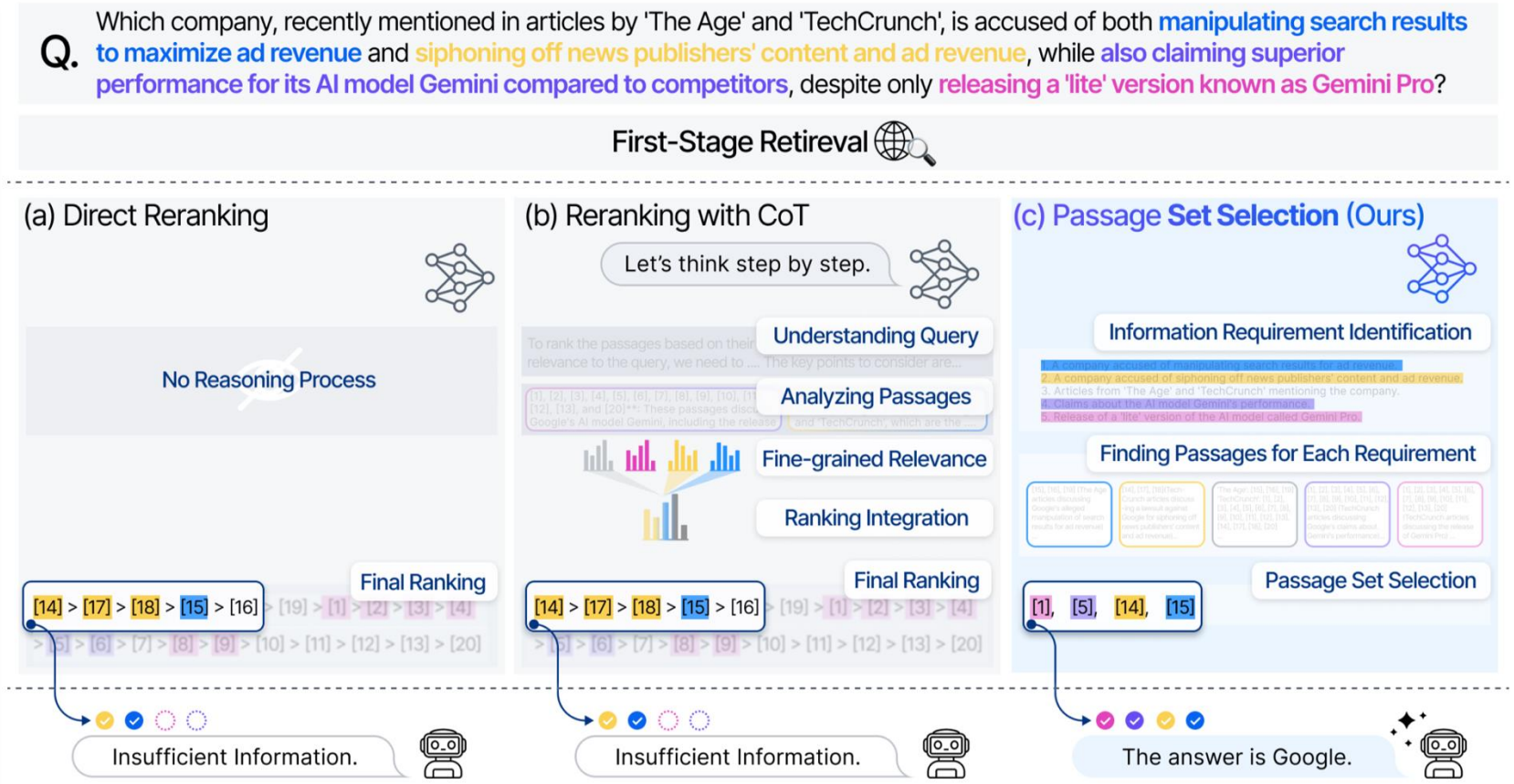


Sun, 2025, Leveraging Outputs of Large Language Model as Feedback for Dynamic Reranking in Retrieval-Augmented Generation. NeurIPS 2025

Retrieved Result Rerank Model

❑ SetR: From reranking to dynamic evidence selection

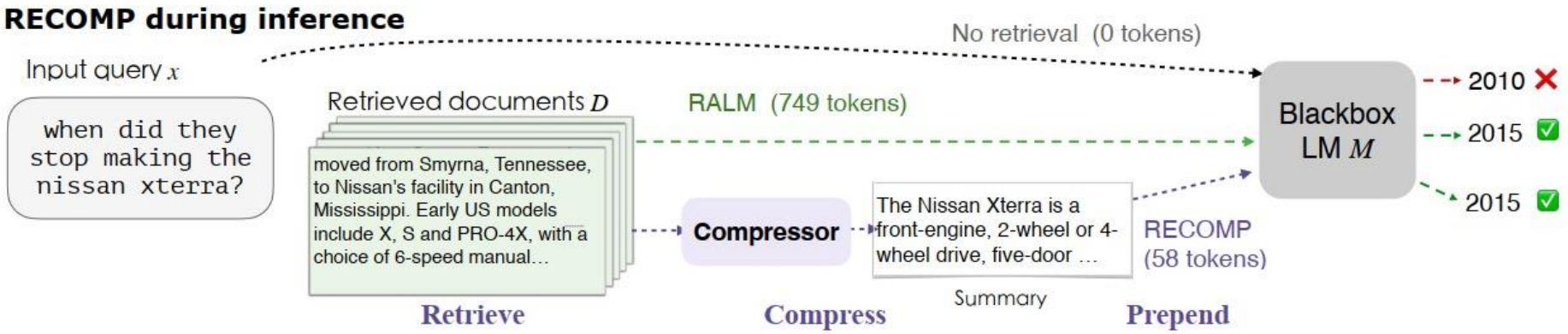
- ❖ **Question**: Are the top-ranked passages the best evidence set?
- ❖ **Problem**: Individual relevance does not guarantee collective sufficiency.
- ❖ **Key idea**: Select a complementary set of passages that covers the query's information needs.



Lee, 2025, Shifting from Ranking to Set Selection for Retrieval Augmented Generation. ACL 2025

Post-Retrieval Techniques

- Retrieved Result Compression
  - To reduce the computational costs and also relieve the burden of LMs to identify relevant information in long retrieved documents.



- Compressor Learning Objectives
  - Concise
  - Effective
  - Faithful

Xu et al. 2024. "RECOMP: Improving retrieval- augmented LMs with context compression and selective augmentation." ICLR 2024

Retrieved Result Compression Performance

- QA tasks

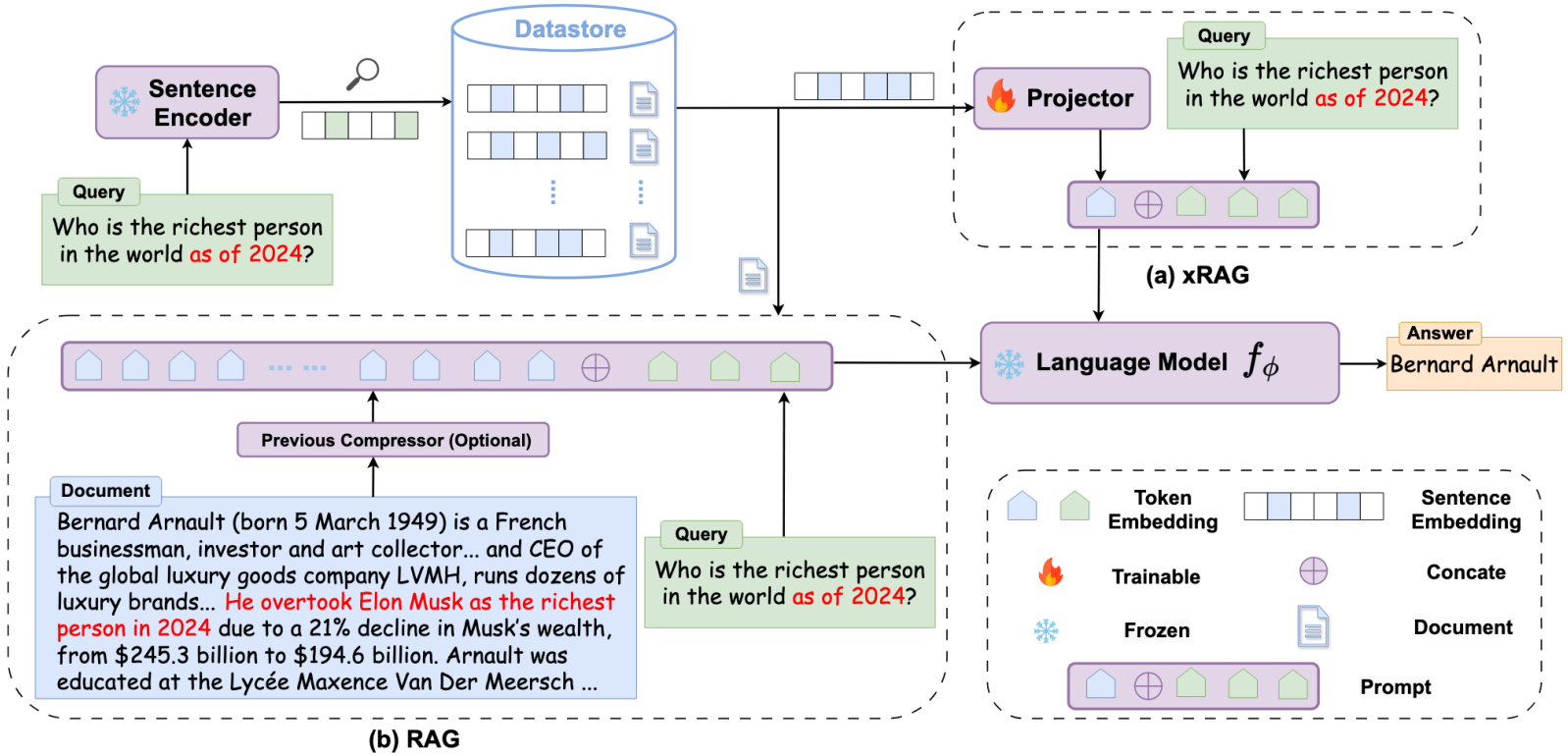
In-Context evidence	NQ # tok	NQ EM	NQ F1	TQA # tok	TQA EM	TQA F1	HotpotQA # tok	HotpotQA EM	HotpotQA F1
-	0	21.99	29.38	0	49.33	54.85	0	17.80	26.10
RALM without compression									
Top 1 documents	132	33.07	41.45	136	57.84	64.94	138	28.80	40.58
Top 5 documents	660	**39.39**	**48.28**	677	**62.37**	**70.09**	684	**32.80**	**43.90**
Phrase/token level compression									
Top 5 documents (NE)	338	23.60	31.02	128	54.96	61.19	157	22.20	31.89
Top 5 documents (BoW)	450	28.48	36.84	259	58.16	65.15	255	25.60	36.00
Extractive compression of top 5 documents									
Oracle	34	60.22	64.25	32	79.29	82.06	70	41.80	51.07
Random	32	23.27	31.09	31	50.18	56.24	61	21.00	29.86
BM25	36	25.82	33.63	37	54.67	61.19	74	26.80	38.02
DPR	39	34.32	43.38	41	56.58	62.96	78	27.40	38.15
Contriever	36	30.06	31.92	40	53.67	60.01	78	28.60	39.48
Ours	37	36.57	44.22	38	**58.99**	65.26	75	**30.40**	**40.14**

Outperforms representative sparse and dense retrievers

Xu et al. 2024. "RECOMP: Improving retrieval- augmented LMs with context compression and selective augmentation." ICLR 2024

Post-Retrieval Techniques

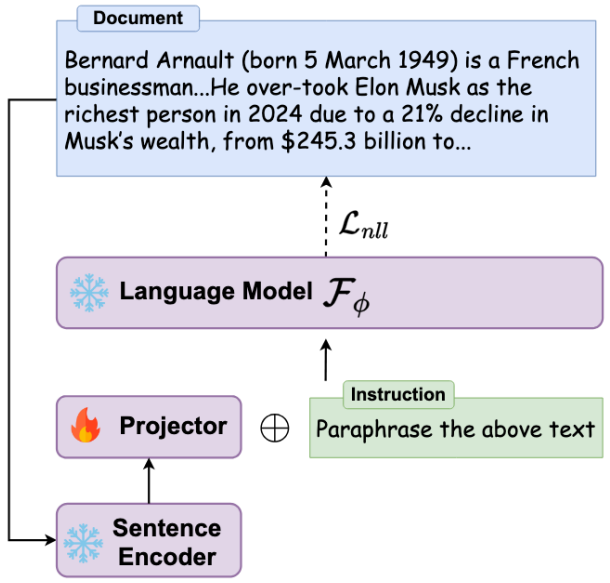
- **xRAG:** Extreme Context Compression
  - xRAG employs a modality fusion methodology to seamlessly integrate document embeddings into the language model's representation space.



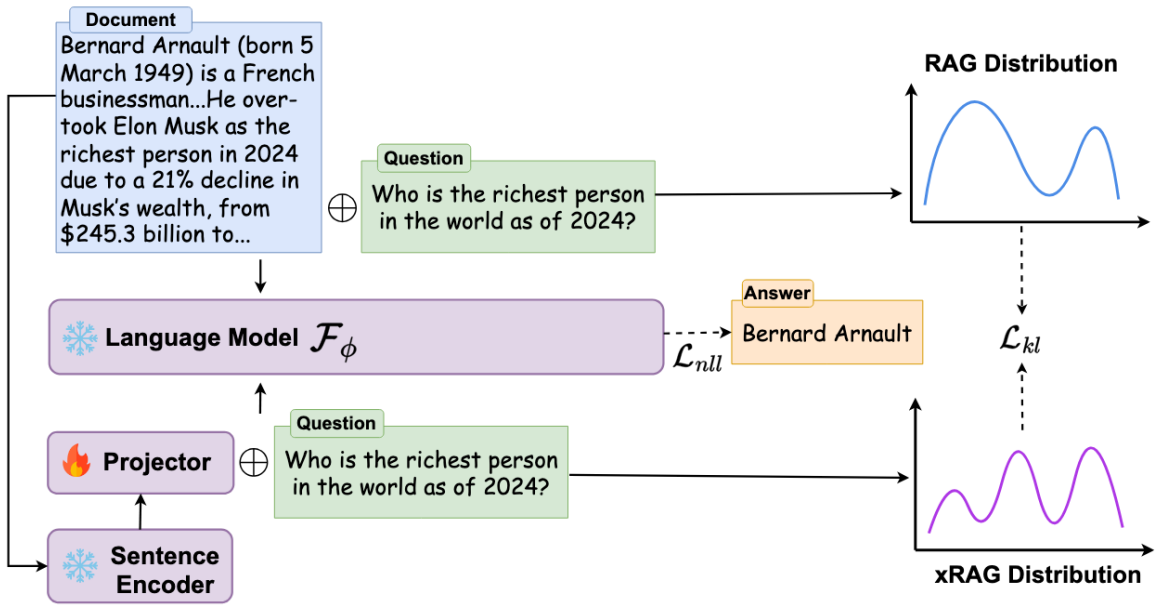
Cheng et al. 2024. "xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token". NeurIPS 2024

Post-Retrieval Techniques

- **xRAG:** Extreme Context Compression



(a) Paraphrase Pretraining



(b) Context-aware Instruction Tuning

$$\mathcal{L}_{\text{nll}} = -\sum_{i=1} \log p_\phi(d_i | \mathbf{W}(\mathrm{E}), \mathbf{X}_{\texttt{instruct}}, d_{<i})$$

$$\mathcal{L}_{\text{nll}} = -\sum_{i=1} \log p_\phi(\mathbf{X}_{\texttt{answer},i} | \mathbf{W}(\mathrm{E}_{\texttt{context}}), \mathbf{X}_{\texttt{question}}, \mathbf{X}_{\texttt{answer},<i})$$

$$\mathcal{L}_{\text{kl}} = D_{\text{KL}}(p_\phi(\mathbf{X}_{\texttt{answer}} | \mathbf{X}_{\texttt{context}}, \cdot) \,||\, p_\phi(\mathbf{X}_{\texttt{answer}} | \mathbf{W}(\mathrm{E}_{\texttt{context}}), \cdot))$$

Cheng et al. 2024. "xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token". NeurIPS 2024

Post-Retrieval Techniques

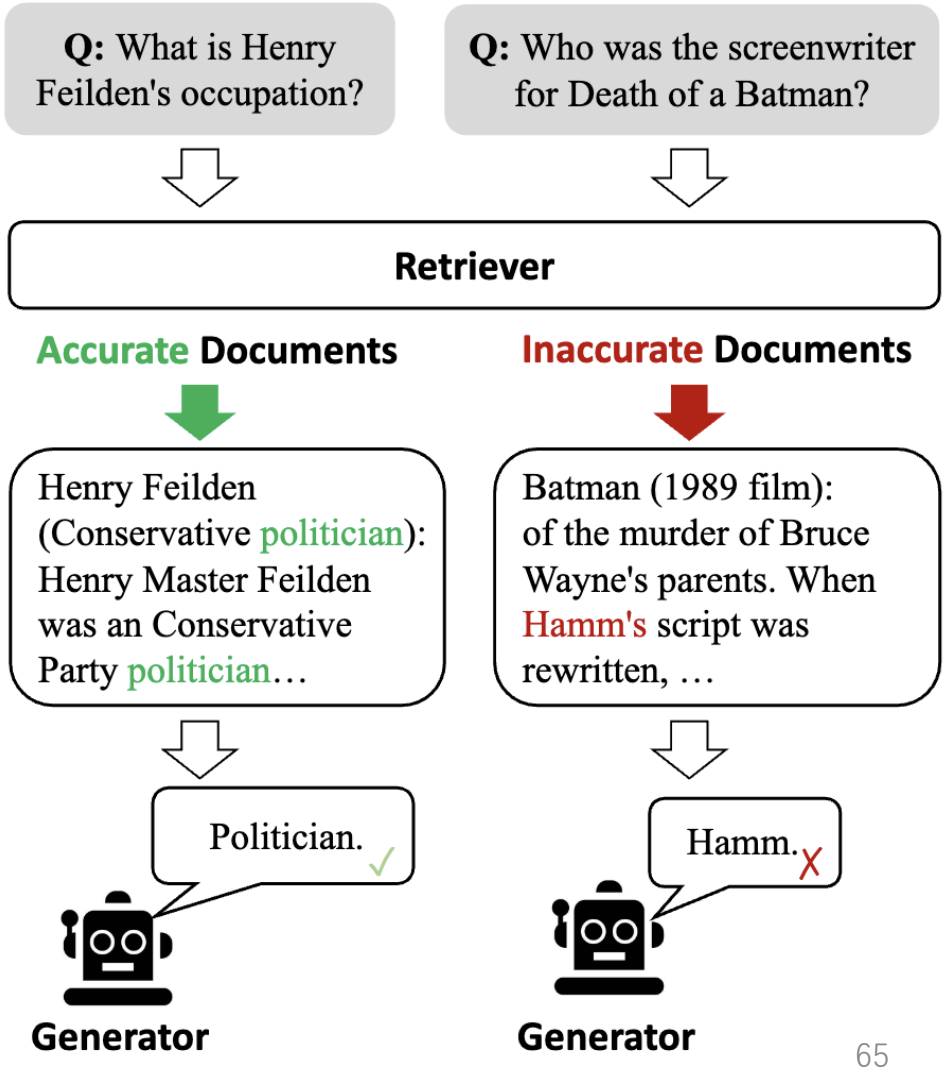
- **xRAG:** Extreme Context Compression

Task Type	NQ	TriviaQA	WebQA	HotpotQA	TrutufulQA		FactKG	Average	# Doc Length
	Open-Domain QA (EM)			Multihop QA (EM)	Long-form QA (F1/R-L)		Fact Checking (Acc)		
Mistral-7b									
w/o retrieval	30.25	57.08	34.89	27.02	26.23	25.51	54.78	36.54 (0.0%)	0
w retrieval	**42.71**	**65.88**	37.84	**38.79**	26.50	25.92	**67.76**	**43.63** (19.4%)	175.1
***with Compression**									
LLMLingua †	30.64	57.94	32.63	29.91	25.70	25.10	64.17	38.01 (4.0%)	98.6
LLMLingua ‡	28.81	57.09	32.33	29.13	26.10	25.39	63.57	37.48 (2.5%)	61.1
TF-IDF	30.25	58.49	35.43	26.62	26.33	25.83	59.56	37.49 (2.6%)	1
xRAG	39.10	65.77	**39.40**	34.05	**28.10**	**27.71**	63.08	42.46 (16.2%)	1
Mixtral-8x7b									
w/o retrieval	41.99	71.10	40.31	32.87	25.60	24.90	62.64	42.76 (0.0%)	0
w retrieval	45.15	70.34	41.26	**43.46**	27.10	25.80	**70.42**	46.22 (8.0%)	175.1
***with Compression**									
LLMLingua†	37.65	67.70	36.02	35.66	25.99	25.39	67.98	42.32 (-1.0%)	96.6
LLMLingua‡	37.81	67.81	35.78	35.27	25.68	25.00	68.03	44.17 (-1.3%)	61.1
TF-IDF	41.19	69.94	41.63	32.05	26.80	26.00	66.17	43.41 (1.4%)	1
xRAG	**47.28**	**74.14**	**44.50**	39.66	**27.80**	**26.64**	68.20	**46.91** (9.7%)	1

Cheng et al. 2024. "xRAG: Extreme Context Compression for Retrieval-augmented Generation with One Token". NeurIPS 2024

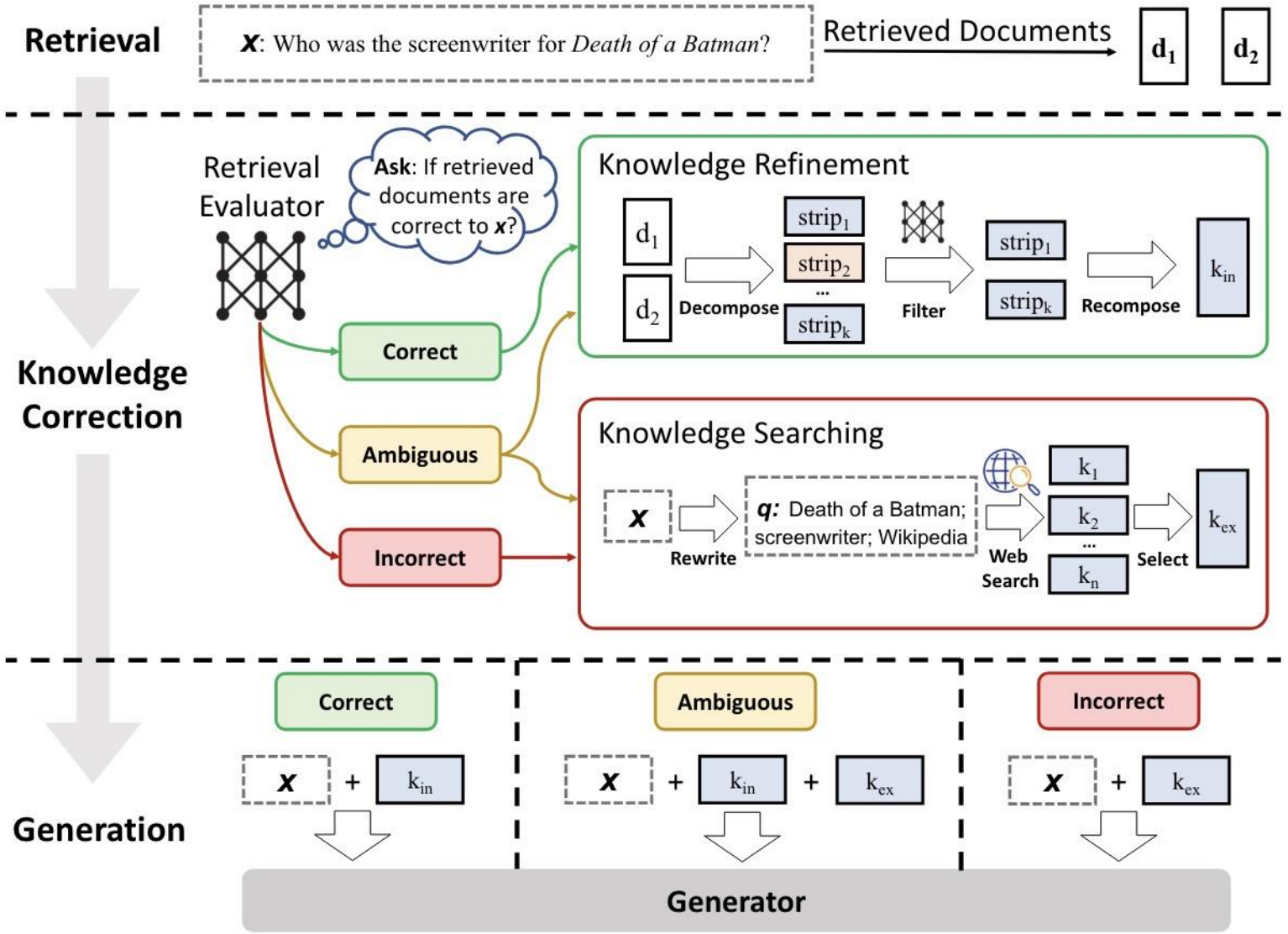
Post-Retrieval Techniques: Corrective RAG

- Corrective Retrieval Augmented Generation (CRAG)
  - Although retrieval-augmented generation (RAG) is a practicable complement to LLMs, it relies heavily on the **relevance of retrieved documents**
  - A lightweight **retrieval evaluator** is designed to assess the overall quality of retrieved documents for a query, returning a **confidence degree** based on which different knowledge retrieval actions can be triggered



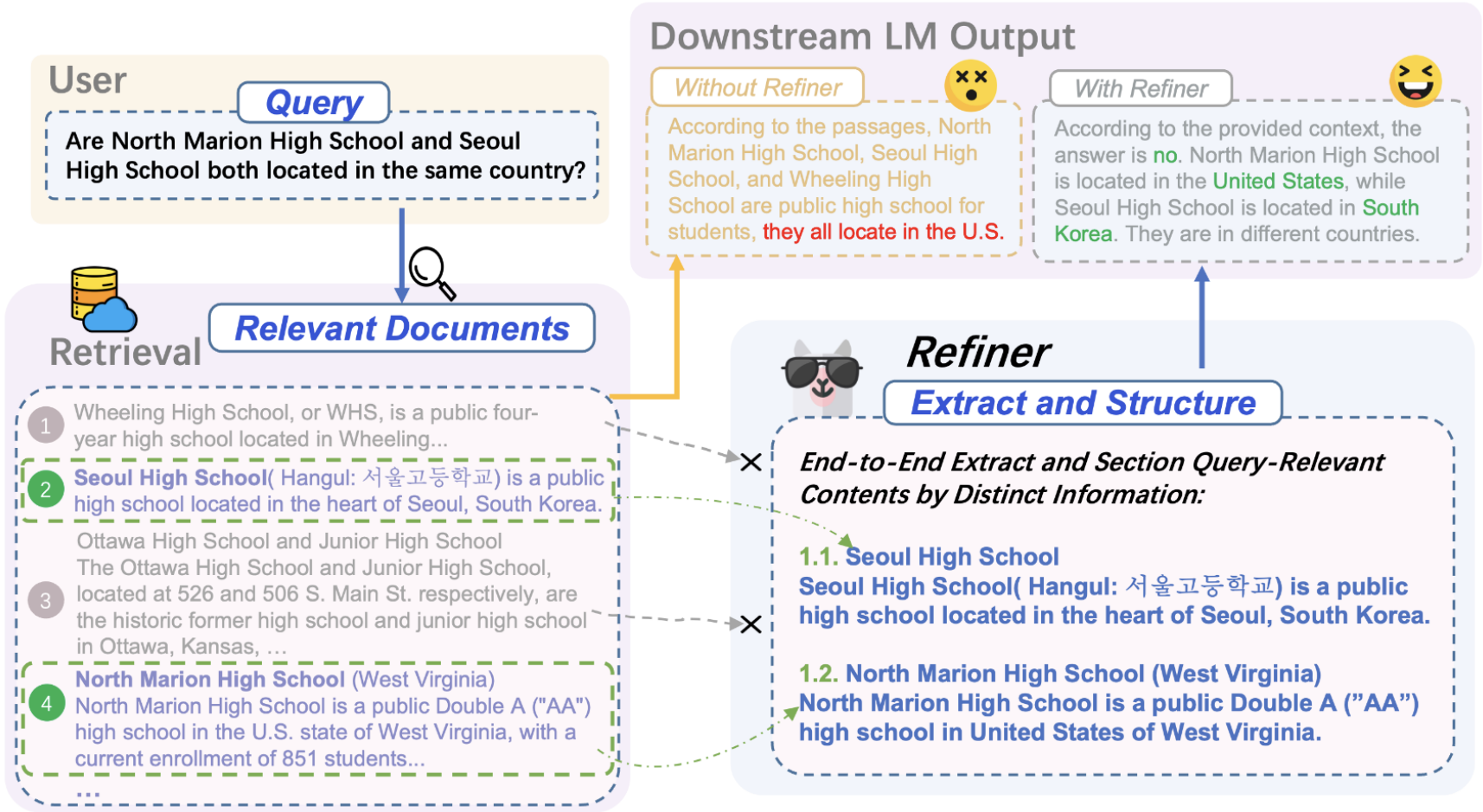
Yan et al., 2024, Corrective Retrieval Augmented Generation

Post-Retrieval Techniques: Corrective RAG



Post-Retrieval Techniques: Refiner

- **Refiner**: leveraging a single decoder-only LLM to adaptively extract query relevant contents verbatim along with the necessary context



Li et al., 2024, "Refiner: Restructure Retrieval Content Efficiently to Advance Question-Answering Capabilities." EMNLP 2024

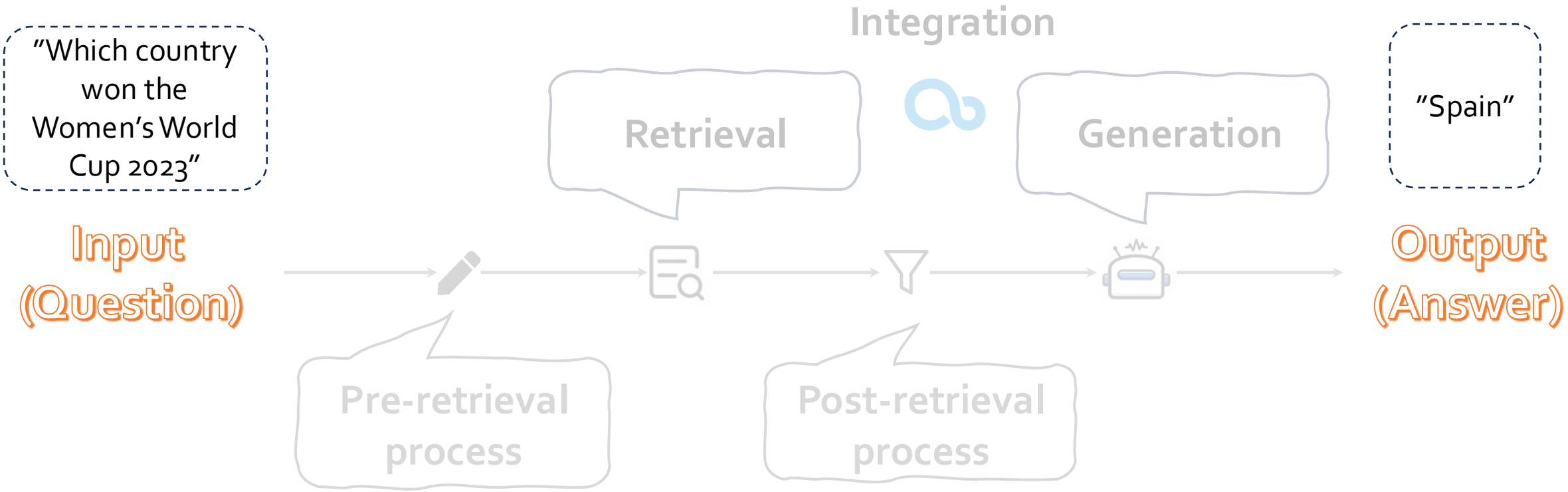
PART 2: Architecture of RA-LFMs and Main Modules


Slides

Website of this tutorial

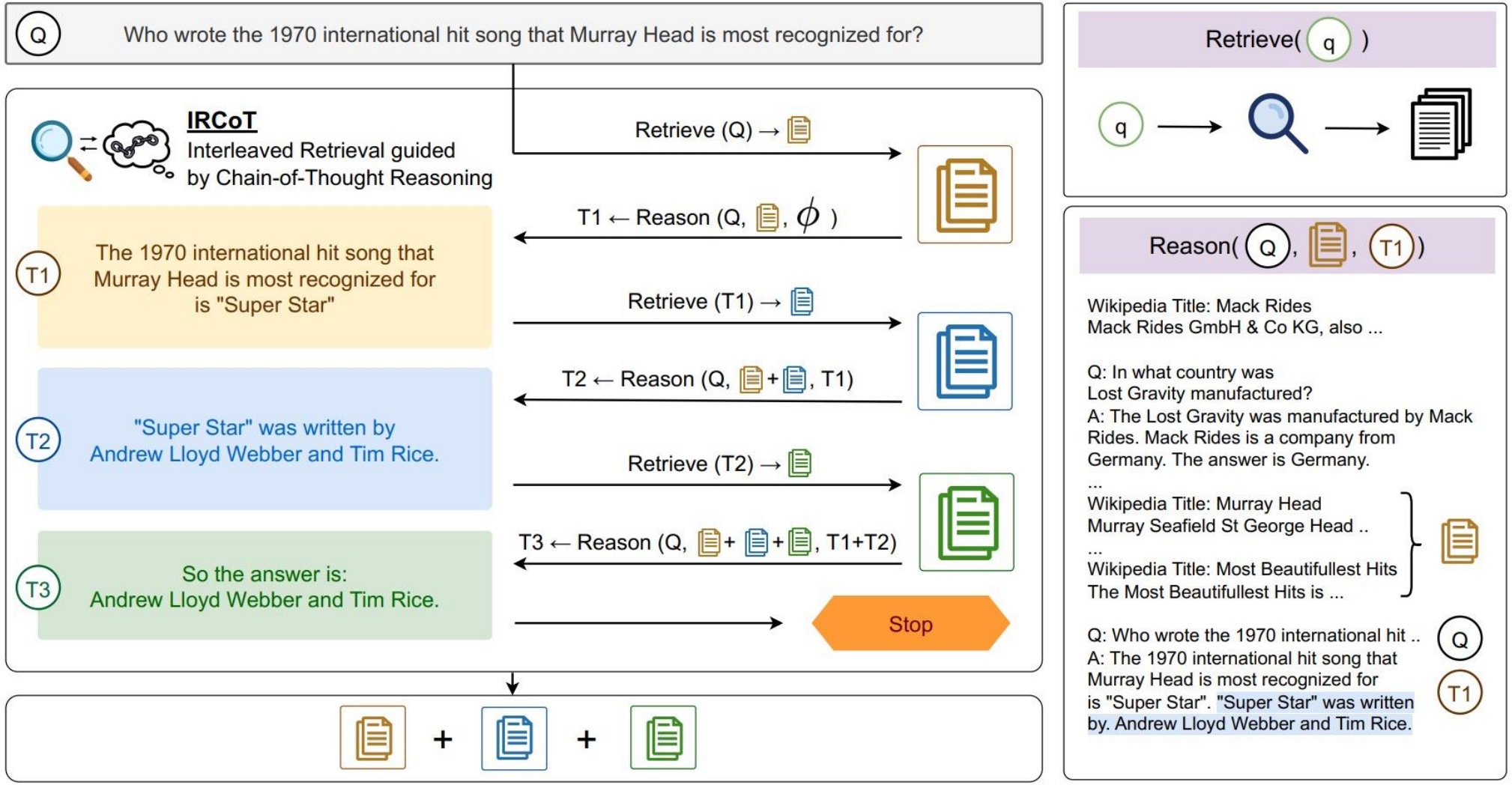
- **RA-LFM architecture overview**
- **Retriever in RA-LFMs**
- **Retrieval results integration**
- **Pre/Post-retrieval techniques**
- **Special RA-LFM paradigms**

Beyond Standard Pipelines and Components?



More?

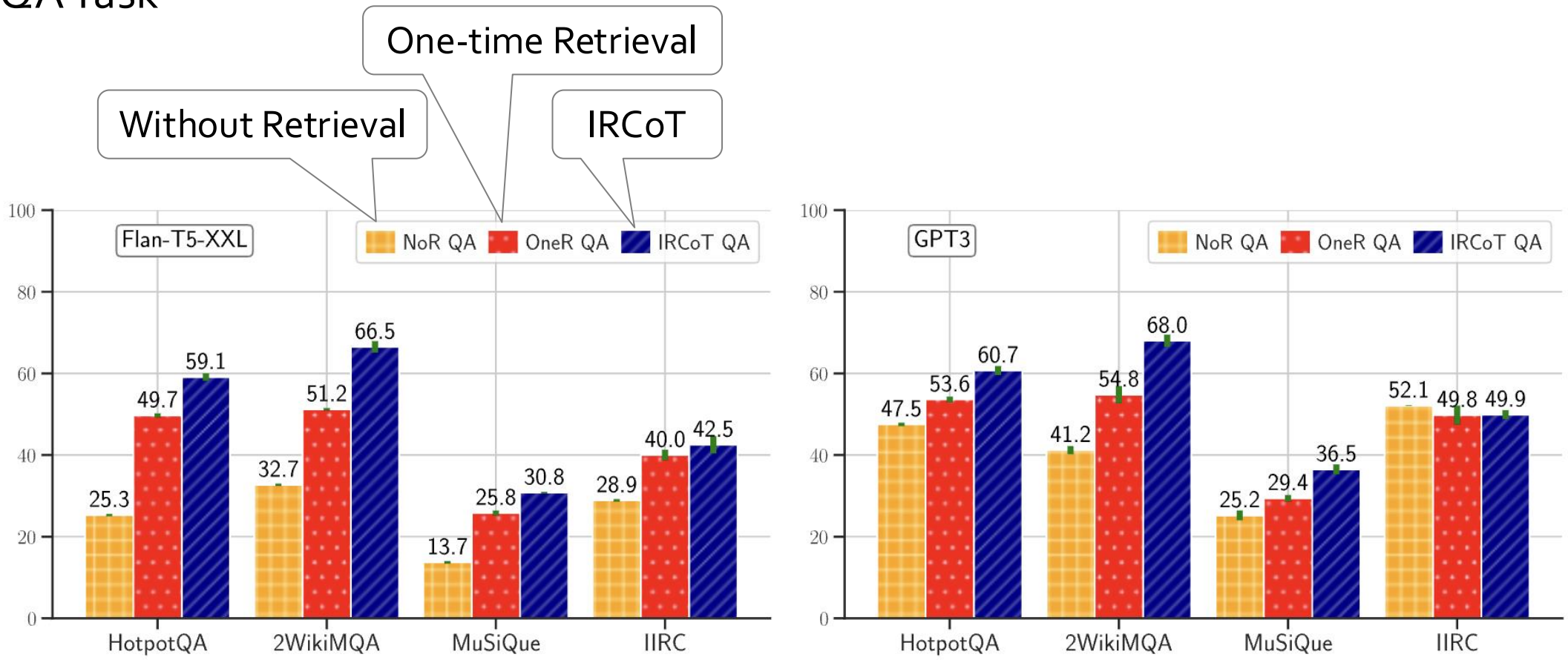
Interleaved Retrieval guided by Chain-of-Thought (IRCoT)



Trivedi, et al. 2023. "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions"

IRCoT Performance

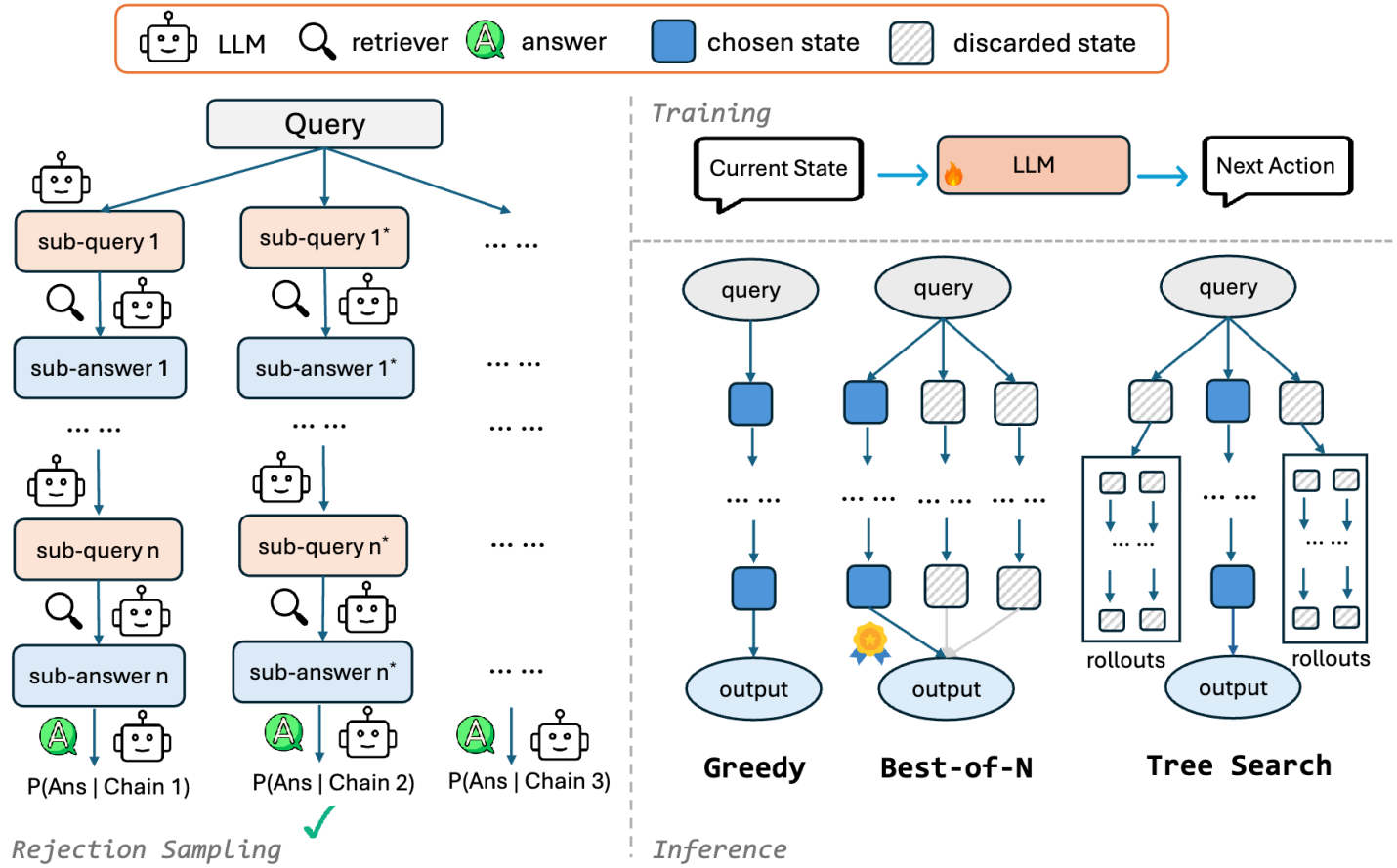
- QA Task



Special RAG Pipeline: Chain-of-RAG

❑ Chain-of-Retrieval Augmented Generation

- ❖ **CoRAG** retrieves and reasons step by step, where each retrieval is conditioned on the evolving reasoning state.
- ❖ **Training**: Rejection sampling builds intermediate retrieval chains.
- ❖ **Inference**: Control chain length and chain number to scale test-time compute.



Wang, et al. 2025. "Chain-of-Retrieval Augmented Generation". NeurIPS 2025

Special RAG Pipeline: Chain-of-RAG

- HotpotQA

Query: What wrestling team is Mathew Thomas Rehwoldt a part of?

RAG without Chain-of-Retrieval
Final Answer: **WWE ✗**

CoRAG (Ours)
Sub-query 1: What is Mathew Thomas Rehwoldt's profession?
Sub-answer 1: No relevant information found.
Sub-query 2: What is Mathew Thomas Rehwoldt's name in the wrestling industry?
Sub-answer 2: Aiden English
Sub-query 3: What wrestling team is Aiden English a part of?
Sub-answer 3: The Vaudevillains
Final Answer: **The Vaudevillains ✓**

Query: How many months apart are Johan Mjällby and Neil Lennon in age?

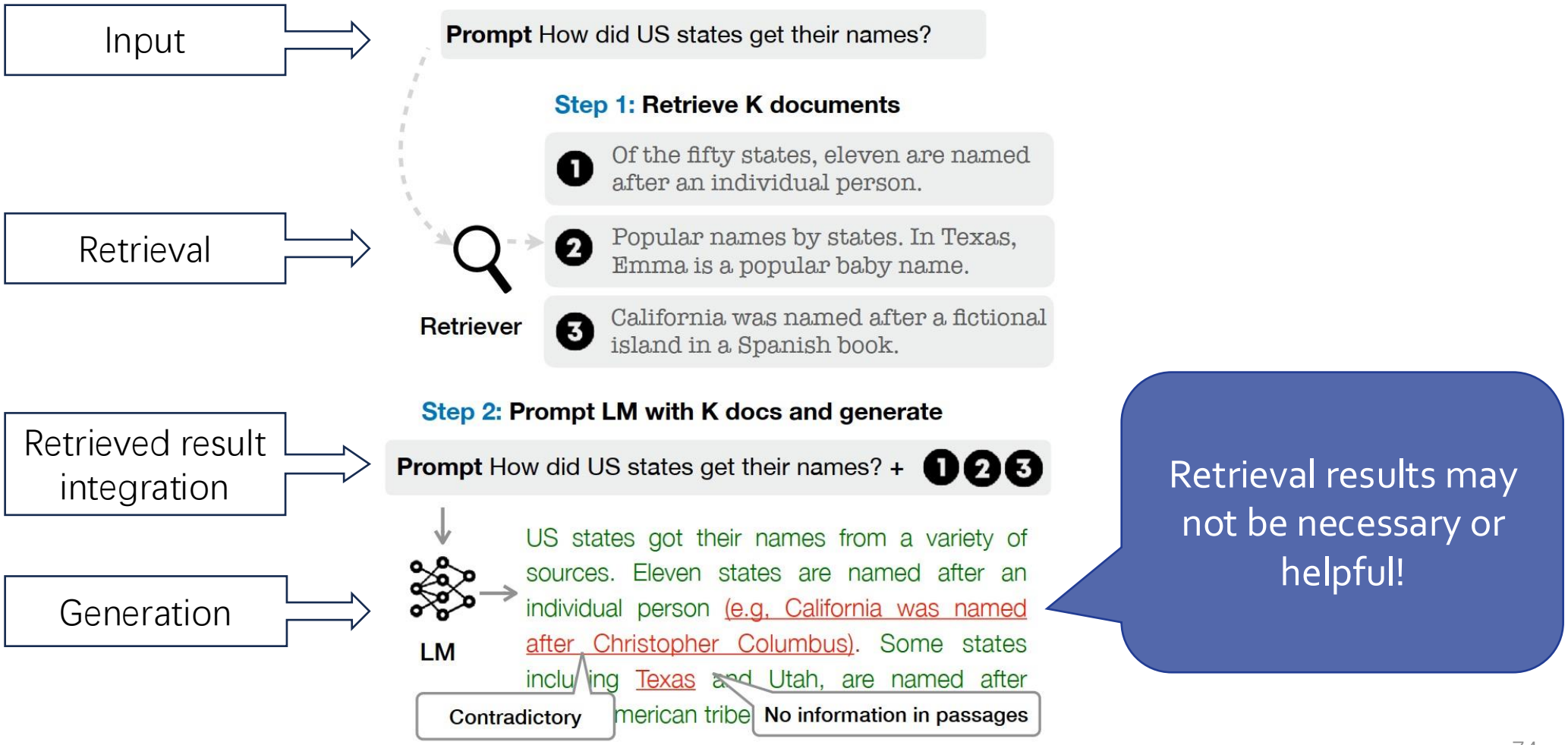
RAG without Chain-of-Retrieval
Final Answer: **two months ✗**

CoRAG (Ours)
Sub-query 1: What is Johan Mjällby's birthdate?
Sub-answer 1: 9 February 1971
Sub-query 2: What is Neil Lennon's birthdate?
Sub-answer 2: 25 June 1971
Sub-query 3: What is the difference in months between 9 February 1971 and 25 June 1971?
Sub-answer 3: 4 months
Final Answer: **4 ✓**

Wang, et al. 2025. "Chain-of-Retrieval Augmented Generation". NeurIPS 2025

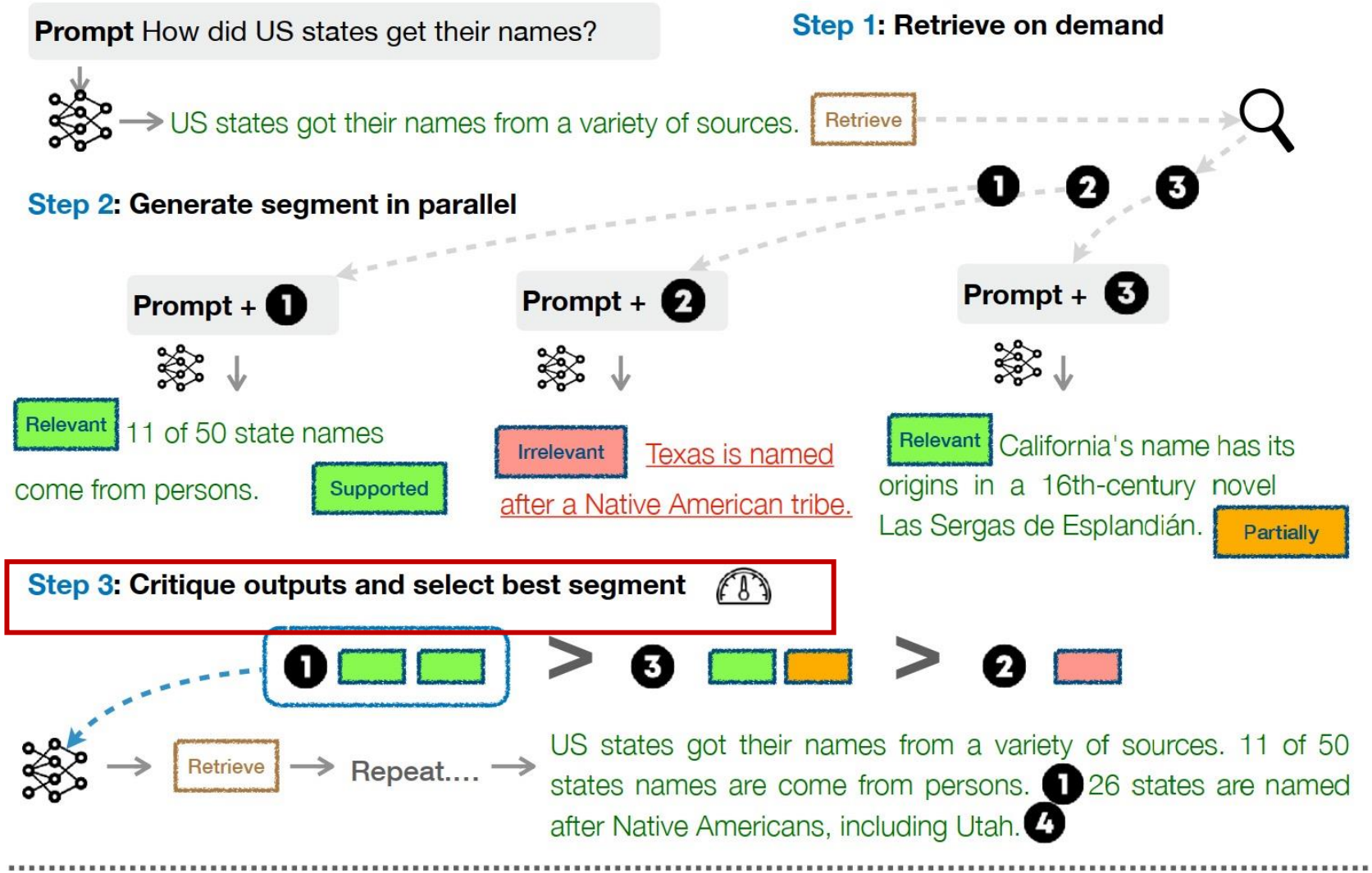
Special RAG Pipeline: Self-Reflective RAG (SELF-RAG)

- General Retrieval-Augmented Generation (RAG)



Asai A, et al. 2024. "Self-rag: Learning to retrieve, generate, and critique through self-reflection." ICLR 2024

SELF-RAG Overview



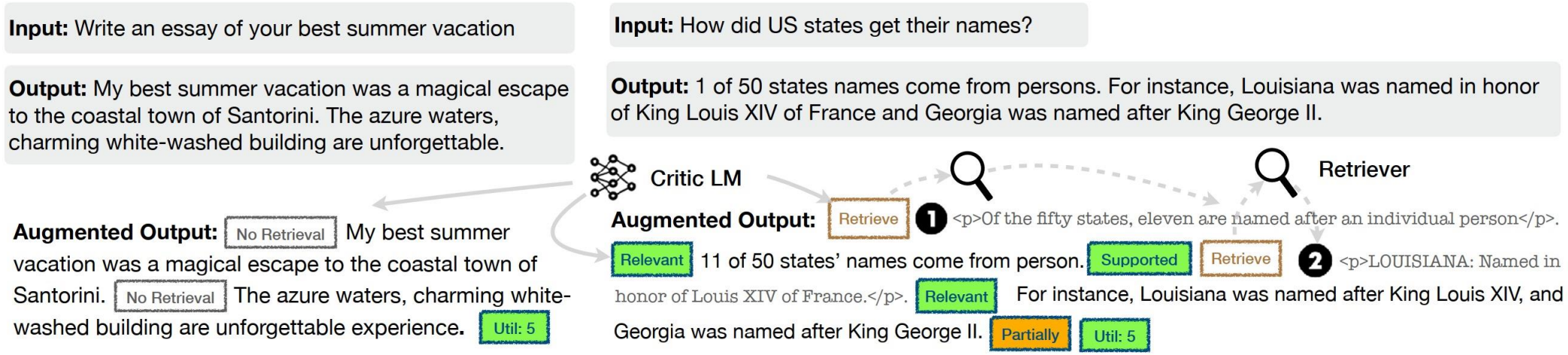
Asai A, et al. 2024. "Self-rag: Learning to retrieve, generate, and critique through self-reflection." ICLR 2024

Key Technical Design in SELF-RAG

- Four types of reflection tokens used in SELF-RAG

Type	Input	Output	Definitions
Retrieve	x / x, y	{yes, no, continue}	Decides when to retrieve with $\mathcal{R}$
ISREL	x, d	{**relevant**, irrelevant}	d provides useful information to solve x.
ISSUP	x, d, y	{**fully supported**, partially supported, no support}	All of the verification-worthy statement in y is supported by d.
ISUSE	x, y	{**5**, 4, 3, 2, 1}	y is a useful response to x.

- Critic Model Training



Asai A, et al. 2024. "Self-rag: Learning to retrieve, generate, and critique through self-reflection." ICLR 2024

SELF-RAG Algorithm

Algorithm 1 SELF-RAG Inference

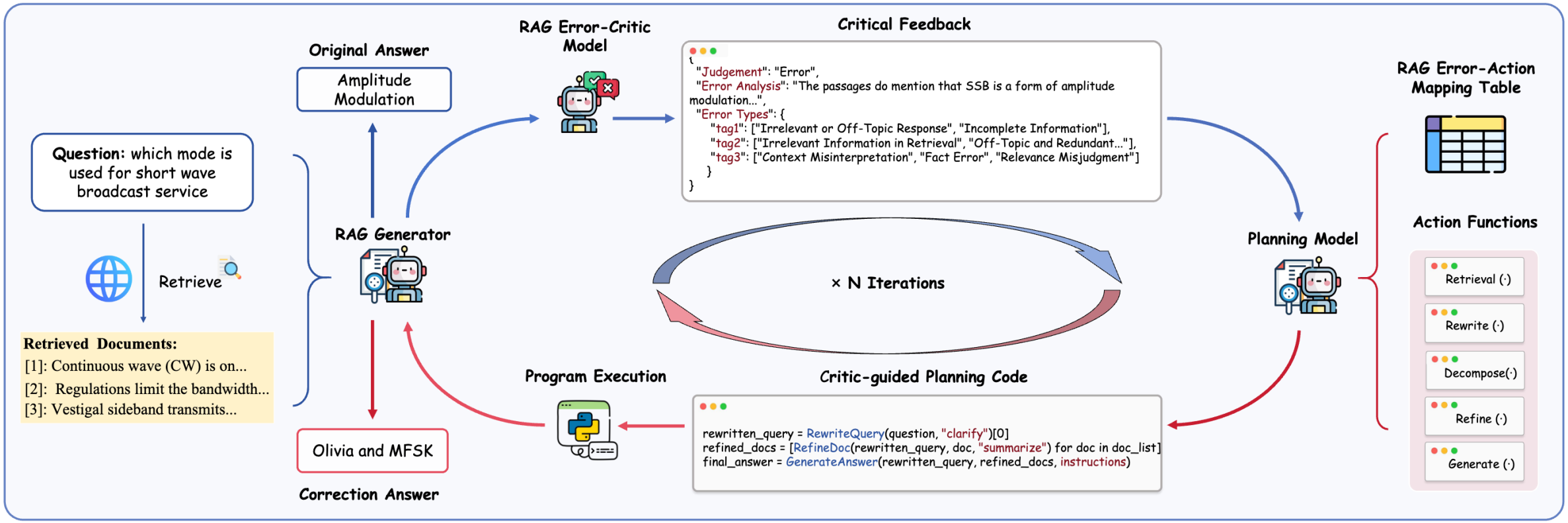
Require: Generator LM $\mathcal{M}$, Retriever $\mathcal{R}$, Large-scale passage collections $\{d_1, \ldots, d_N\}$

Input: input prompt x and preceding generation $y_{<t}$, **Output:** next output segment y_t
$\mathcal{M}$ predicts Retrieve given $(x, y_{<t})$
if Retrieve == Yes **then**
    Retrieve relevant text passages **D** using $\mathcal{R}$ given (x, y_{t-1}) ▷ Retrieve
    $\mathcal{M}$ predicts IsRel given x, d and y_t given $x, d, y_{<t}$ for each $d \in \mathbf{D}$ ▷ Generate
    $\mathcal{M}$ predicts IsSup and IsUse given x, y_t, d for each $d \in \mathbf{D}$ ▷ Critique
    Rank y_t based on IsRel , IsSup , IsUse ▷ Detailed in Section 3.3
else if Retrieve == No **then**
    $\mathcal{M}_{gen}$ predicts y_t given x ▷ Generate
    $\mathcal{M}_{gen}$ predicts IsUse given x, y_t ▷ Critique

Asai A, et al. 2024. “Self-rag: Learning to retrieve, generate, and critique through self-reflection.” ICLR 2024

Modern extension: Critic-guided RAG

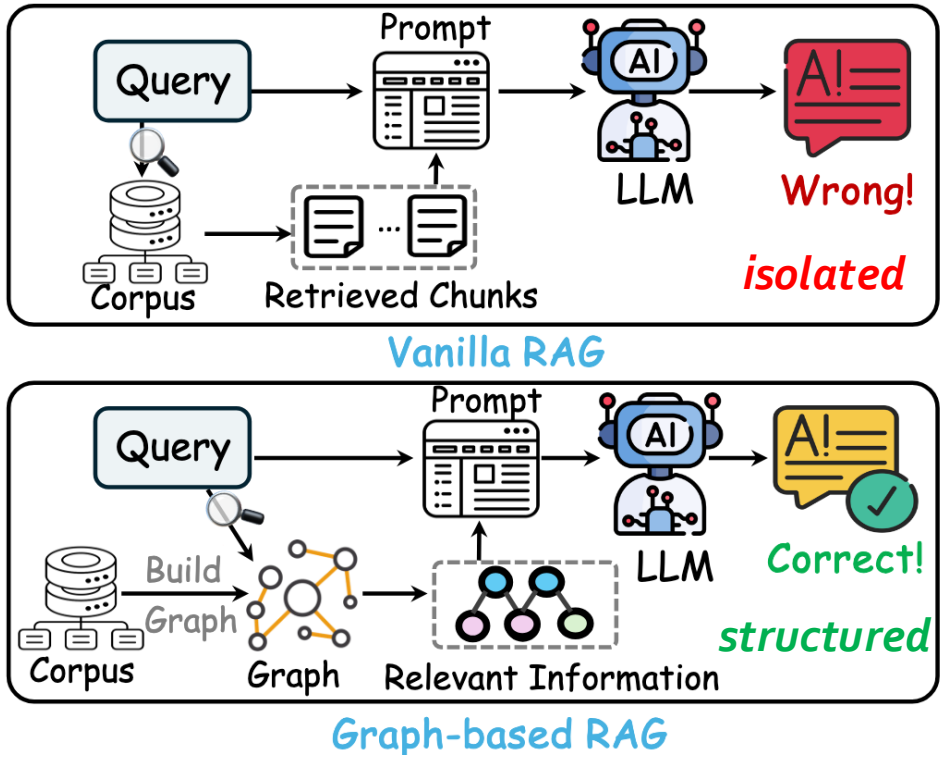
- **RAG-Critic**: a critic-guided agentic RAG framework, enabling the planning agent to customize executor-based solution programs based on the error-critic model's feedback, facilitating an automated error-driven self-correction process.

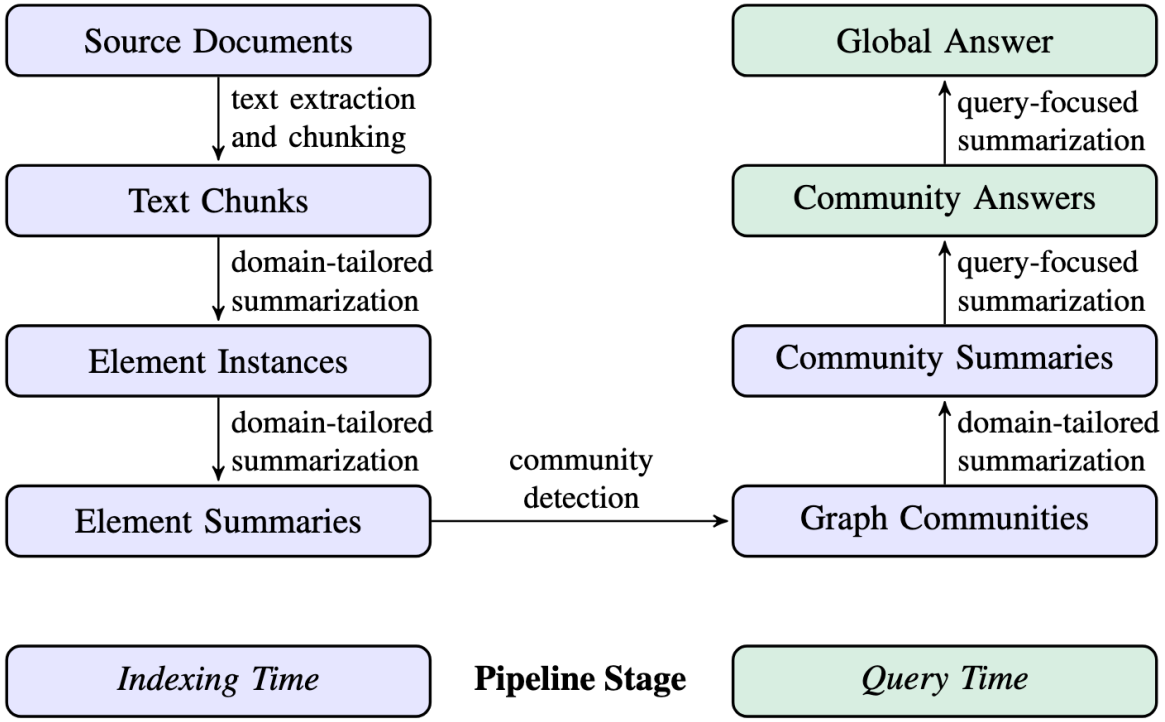


Dong, et al., 2025, RAG-Critic: Leveraging Automated Critic-Guided Agentic Workflow for Retrieval Augmented Generation. ACL 2025

Graph Retrieval-Augmented Generation

- **GraphRAG** retrieves structured evidence over entities, relations, and communities. It is better suited for global, multi-hop, and relational questions.

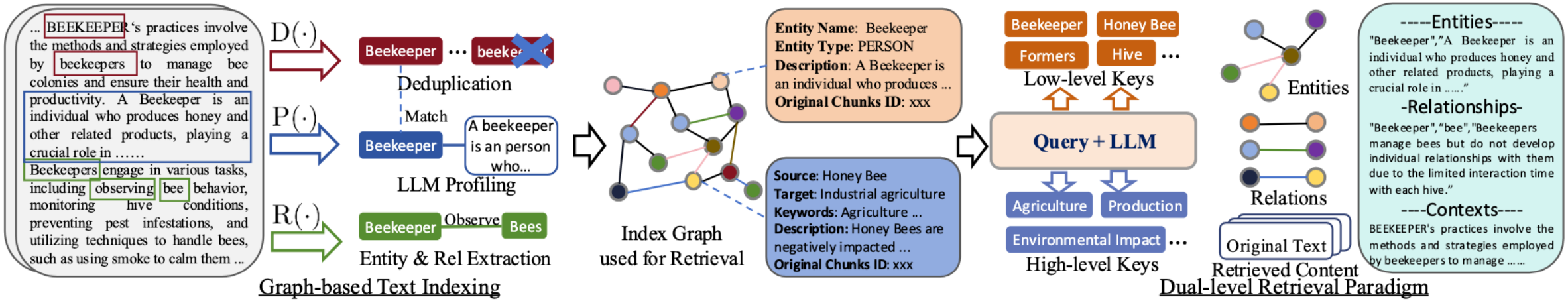




Edge, et al. 2024. "From Local to Global: A Graph RAG Approach to Query-Focused Summarization".

GraphRAG for Lightweight and Incremental Retrieval

❑ **LightRAG:** efficient and incremental GraphRAG through dual-level graph retrieval



- ❖ Graph structure improves evidence organization.
- ❖ Vector retrieval improves efficiency.
- ❖ Incremental updates reduce rebuilding cost.

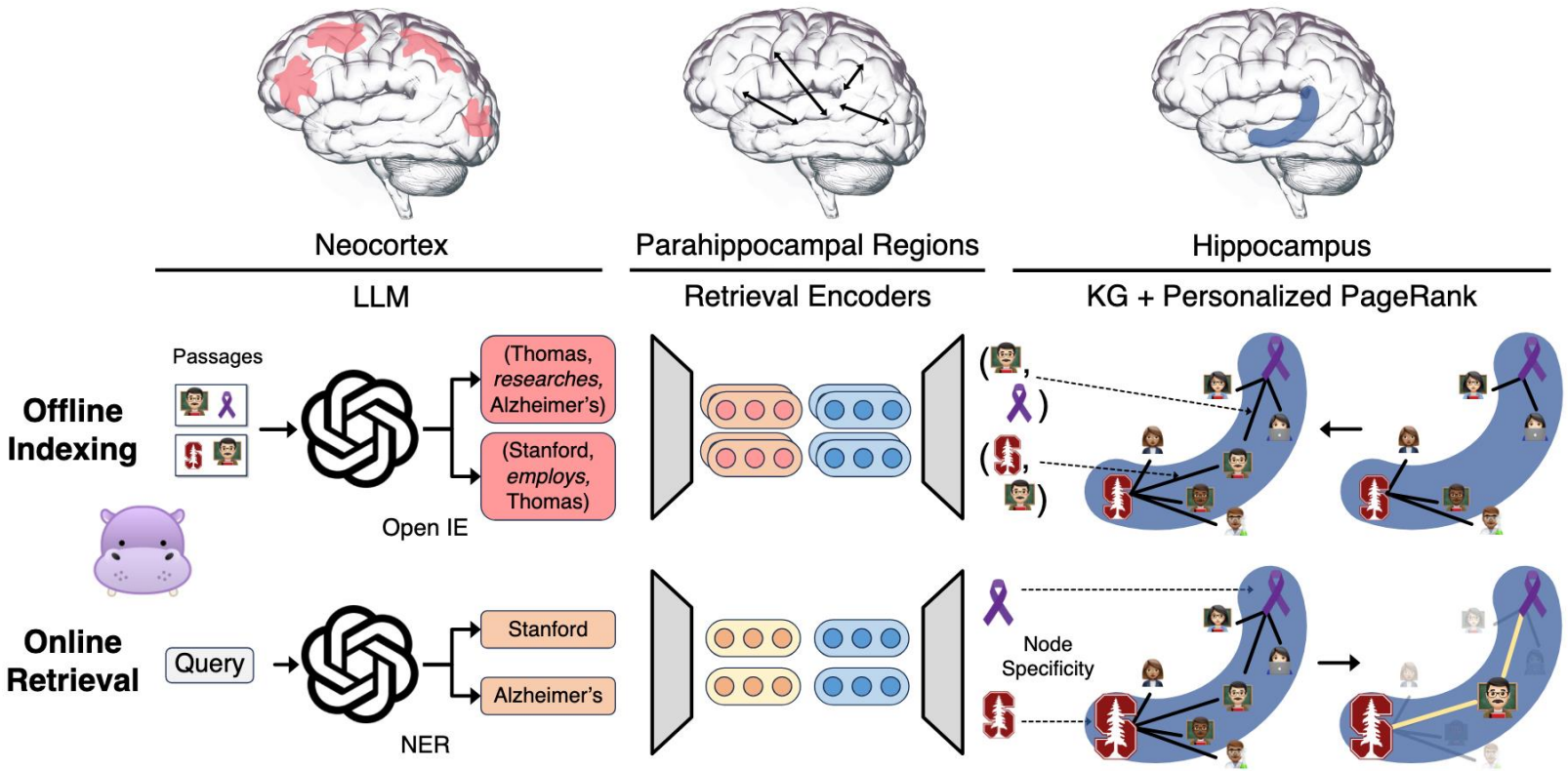
Phase	Retrieval Phase		Incremental Text Update	
Model	GraphRAG	Ours	GraphRAG	Ours
Tokens	$610 \times 1{,}000$	< 100	$1{,}399 \times 2 \times 5{,}000 + T_{\text{extract}}$	T_{extract}
API Calls	$\frac{610 \times 1{,}000}{C_{\max}}$	1	$1{,}399 \times 2 + C_{\text{extract}}$	C_{extract}

Guo, et al. 2025. "LightRAG: Simple and Fast Retrieval-Augmented Generation". EMNLP 2025

GraphRAG as Associative Long-Term Memory

- **HippoRAG:** Using graph memory and associative activation for multi-hop knowledge retrieval

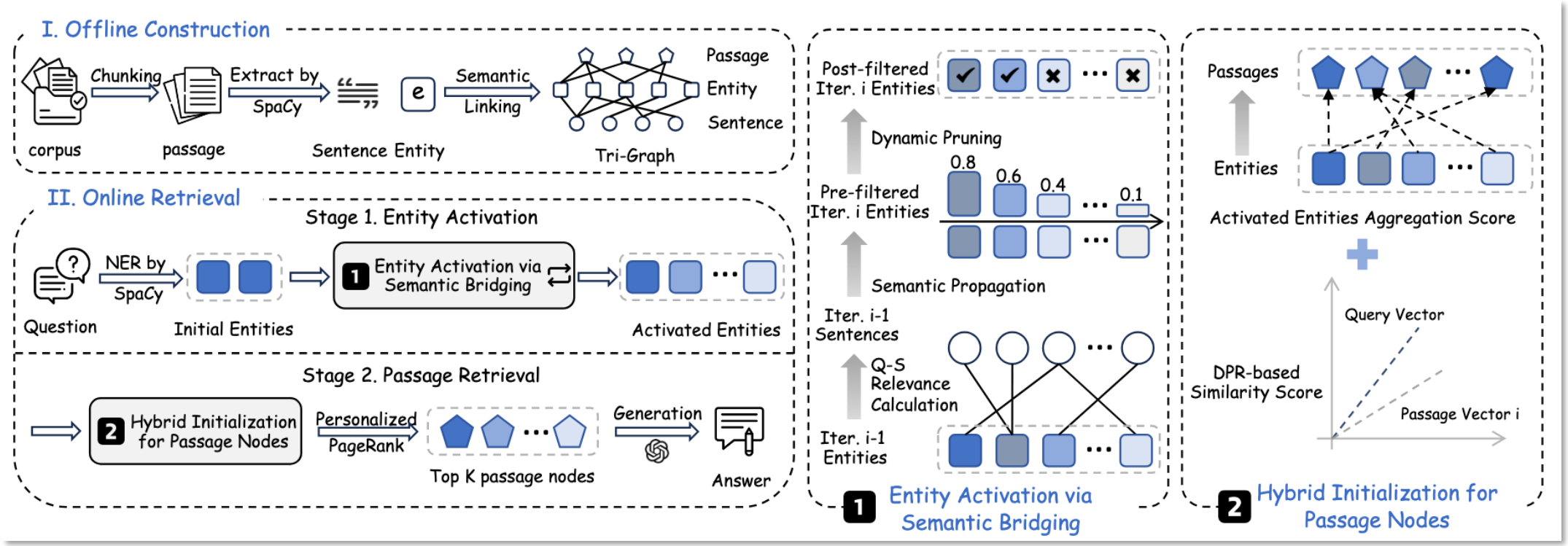
- Knowledge graph as associative memory
- Personalized PageRank retrieval
- Efficient multi-hop knowledge integration



Gutiérrez, et al. 2024. "HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models". NeurIPS 2024

GraphRAG for Scalable Graph Construction

❑ **LinearRAG:** Scaling GraphRAG by avoiding costly relation extraction on large corpora



❖ Relation-free hierarchical graph

❖ Entity activation

❖ Global importance aggregation

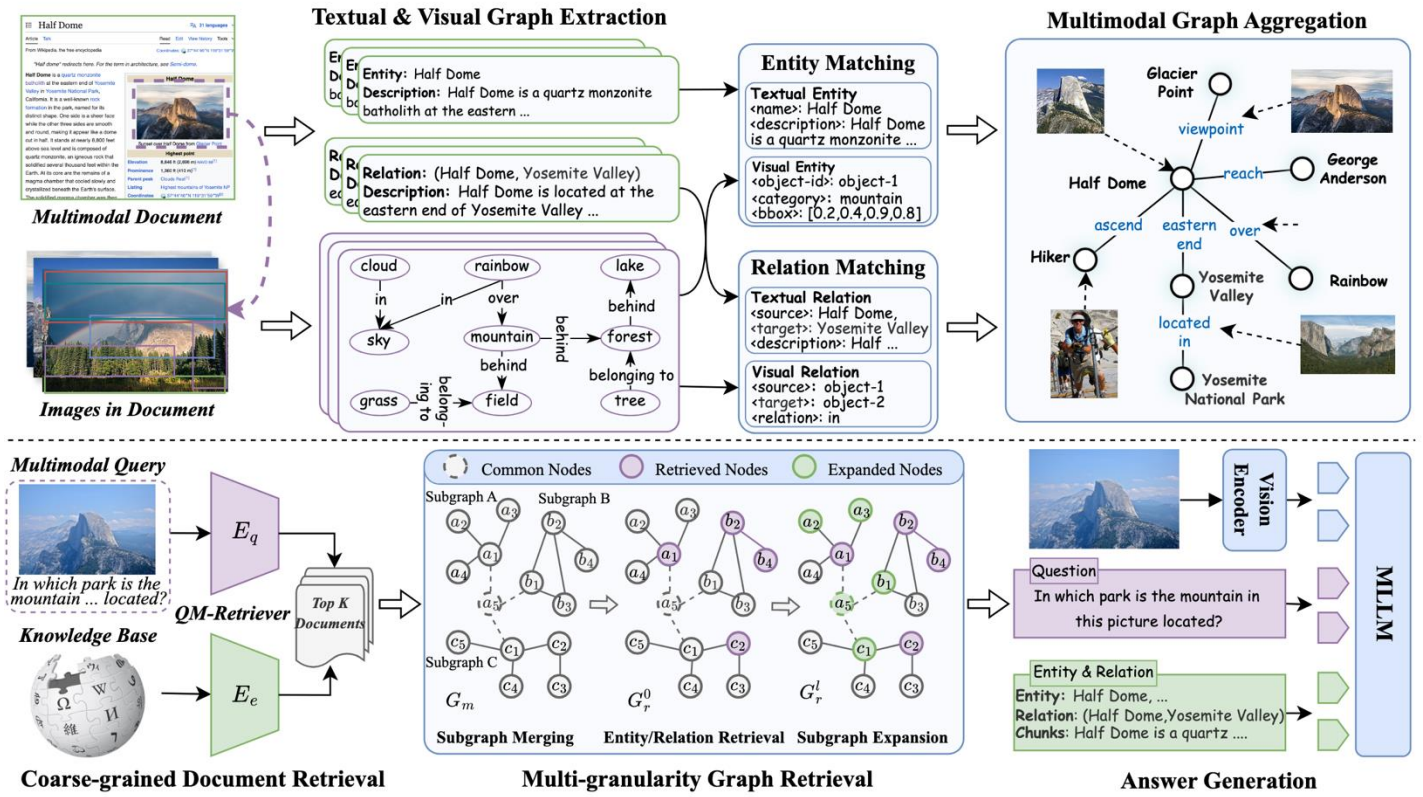
❖ Linear-scale graph construction

Zhuang, et al. 2025. "LinearRAG: Linear Graph Retrieval Augmented Generation on Large-scale Corpora". ICLR 2026

Graph RAG for Multimodal Corpus

❑ **mKG-RAG:** Extending GraphRAG to multimodal knowledge graphs

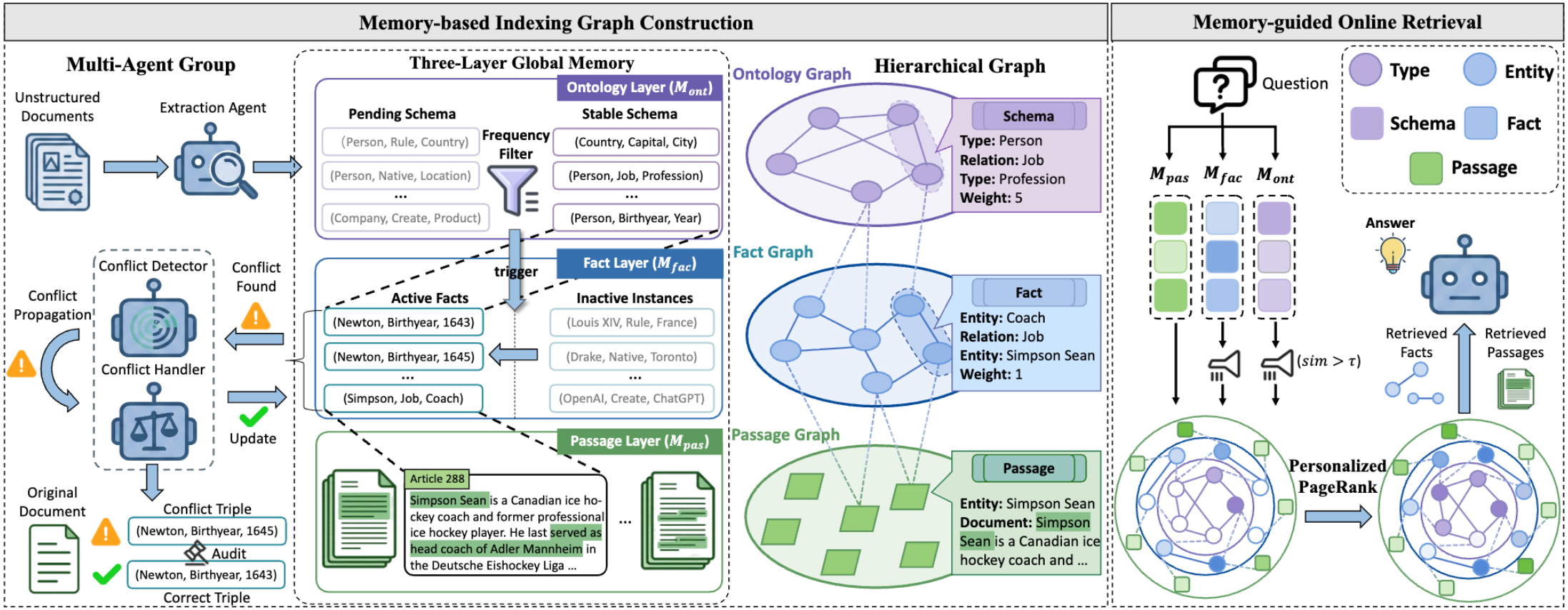
- ❖ Knowledge-intensive VQA requires cross-modal evidence.
- ❖ Multimodal KGs align textual and visual entities and relationships
- ❖ Multi-granularity Graph retrieval provides structured multimodal context.



Yuan, et al. 2025. "Leveraging Multimodal Knowledge Graphs in Retrieval-Augmented Generation for Knowledge-intensive VQA". SIGIR 2026

GraphRAG with Memory-Based Multi-Agent Construction

- **MemGraphRAG:** Building coherent corpus-level graphs with memory-based multi-agent construction



Wu, et al. 2025. "MemGraphRAG: Memory-based Multi-Agent System for Graph Retrieval-Augmented Generation". KDD 2026

Tutorial Outline

- **Part 1: Introduction** of Retrieval Augmented Large Foundation Models (RA-LFMs) (Dr. Wenqi Fan)
- **Part 2: Architecture** of RA-LFMs and Main Modules (Xu Yuan)
- **Part 3: Learning Approach of RA-LFMs (Chengliang Liu)**
- **Part 4: Agentic RAG** (Chengliang Liu)
- **Part 5: Applications** of RA-LFMs (Chun-Hin Chan)
- **Part 6: Challenges and Future Directions** of RA-LFMs (Dr. Wenqi Fan)
- **Part 7: Q&A**

Website of this tutorial
Check out the slides and more information!

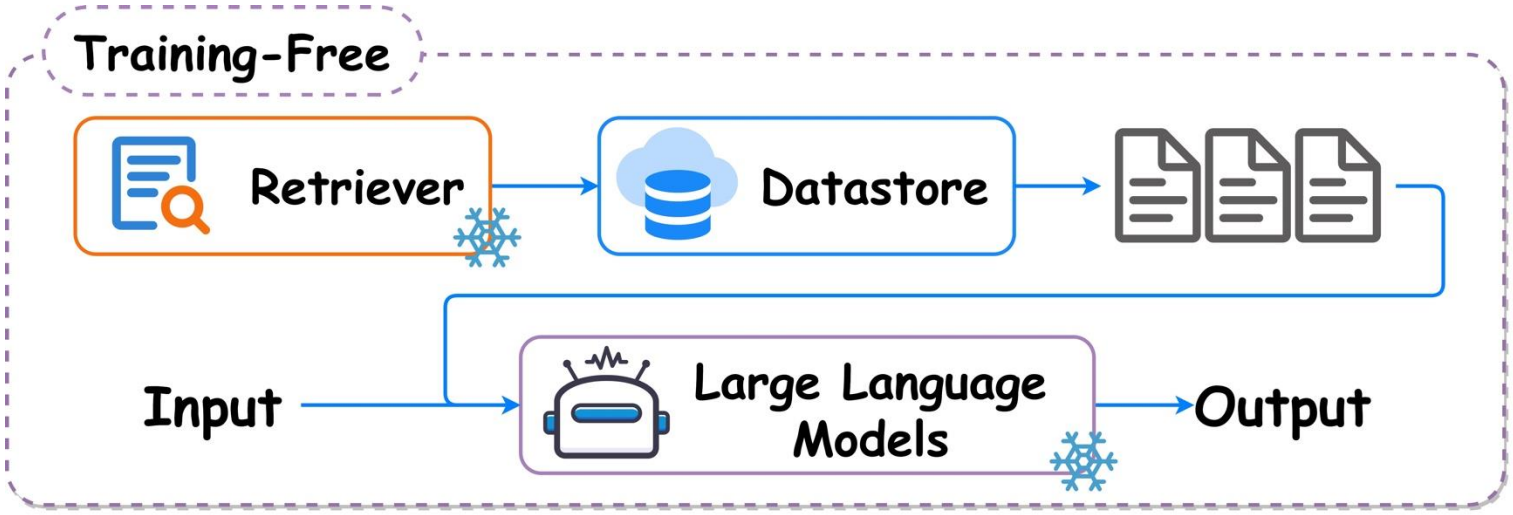
Part 3: RA-LFM Learning

Presenter
Chengliang Liu
HK PolyU

- **Training-free Methods**
- Training-based Methods
  - Independent Learning
  - Sequential Learning
  - Joint Learning

RA-LFM Learning: Training-free

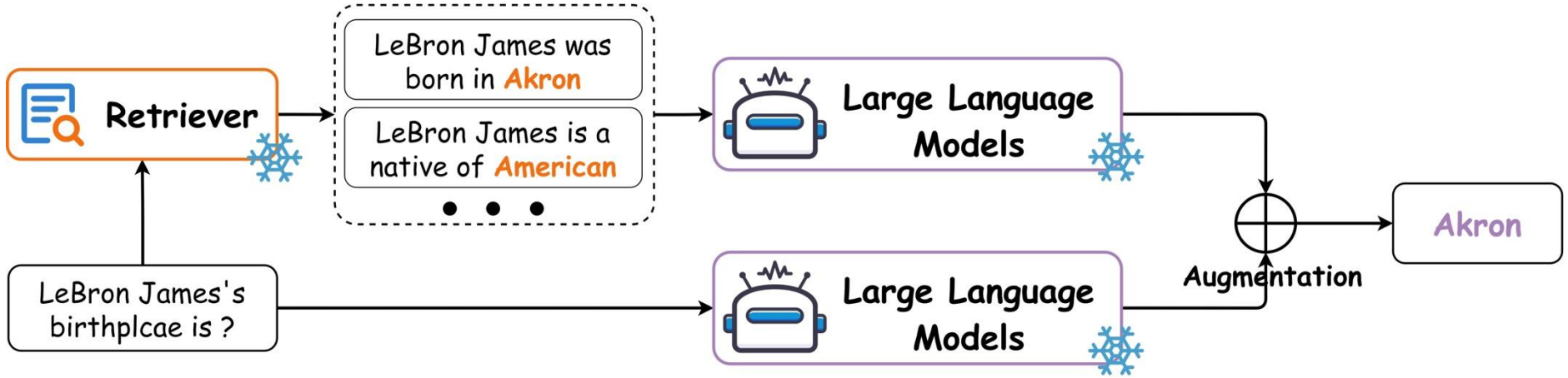
- **Retrieval models** and **language models** are both **frozen**.



RA-LFM Learning: Training-free

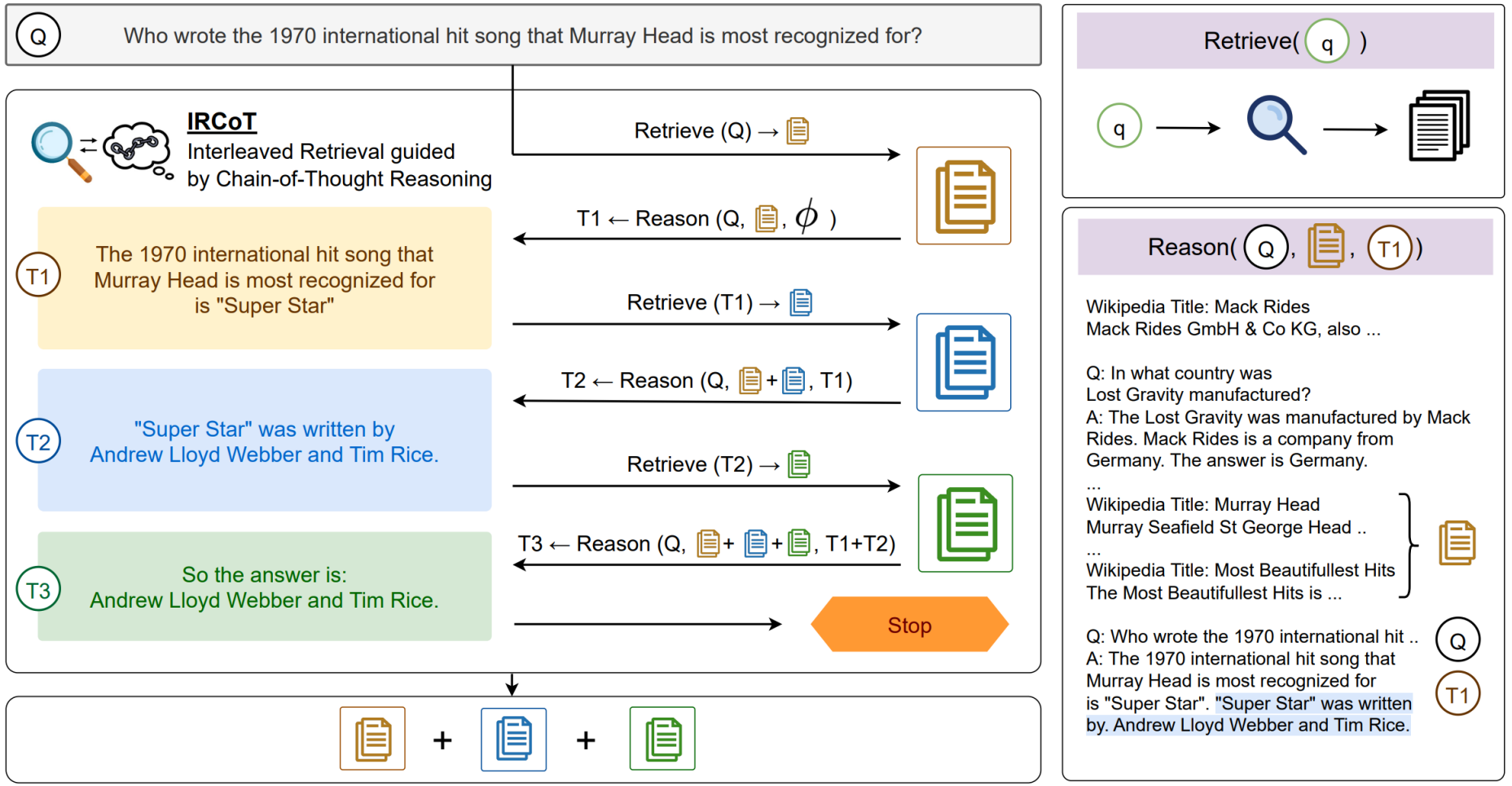
- **Prompt Engineering-based Methods**

- **Retrieval-Guided Token Generation Methods**



RA-LFM Learning: Training-free

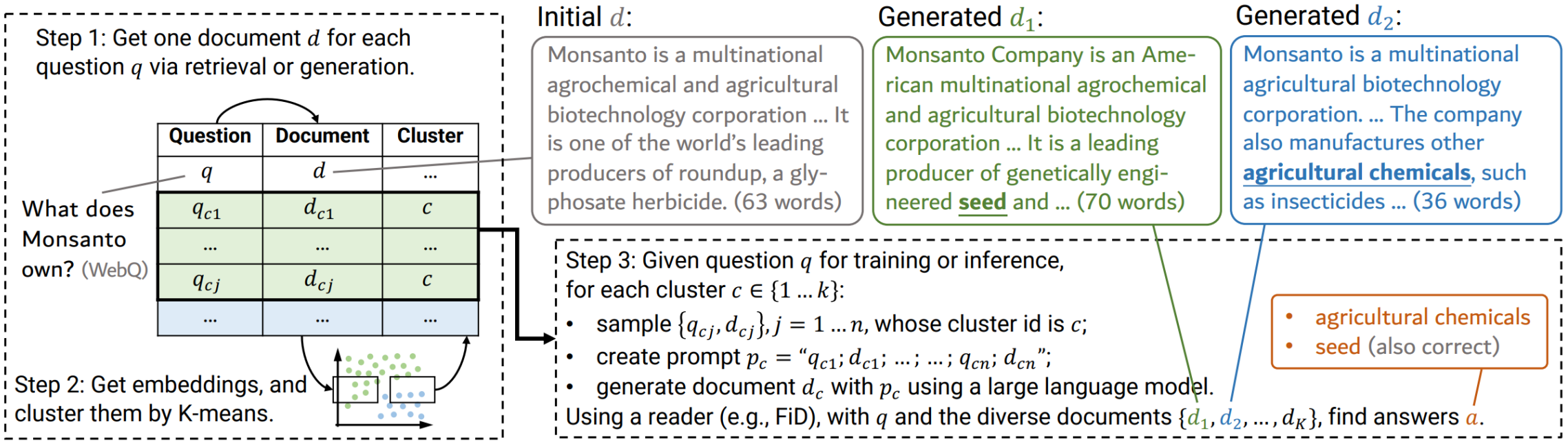
- **IRCoT**



Trivedi, Harsh, et al. "Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions." ACL. 2023.

RA-LFM Learning: Training-free

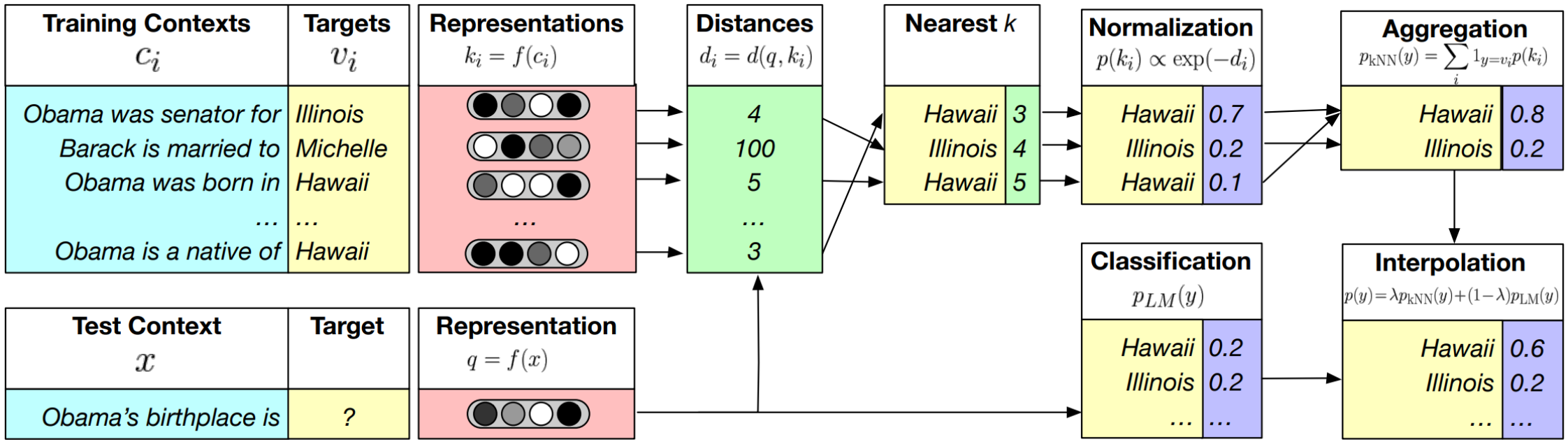
- **GENREAD**



Yu, Wenhao, et al. "Generate rather than Retrieve: Large Language Models are Strong Context Generators." International Conference on Learning Representations. 2023.

RA-LFM Learning: Training-free

- ***k*NN-LM**

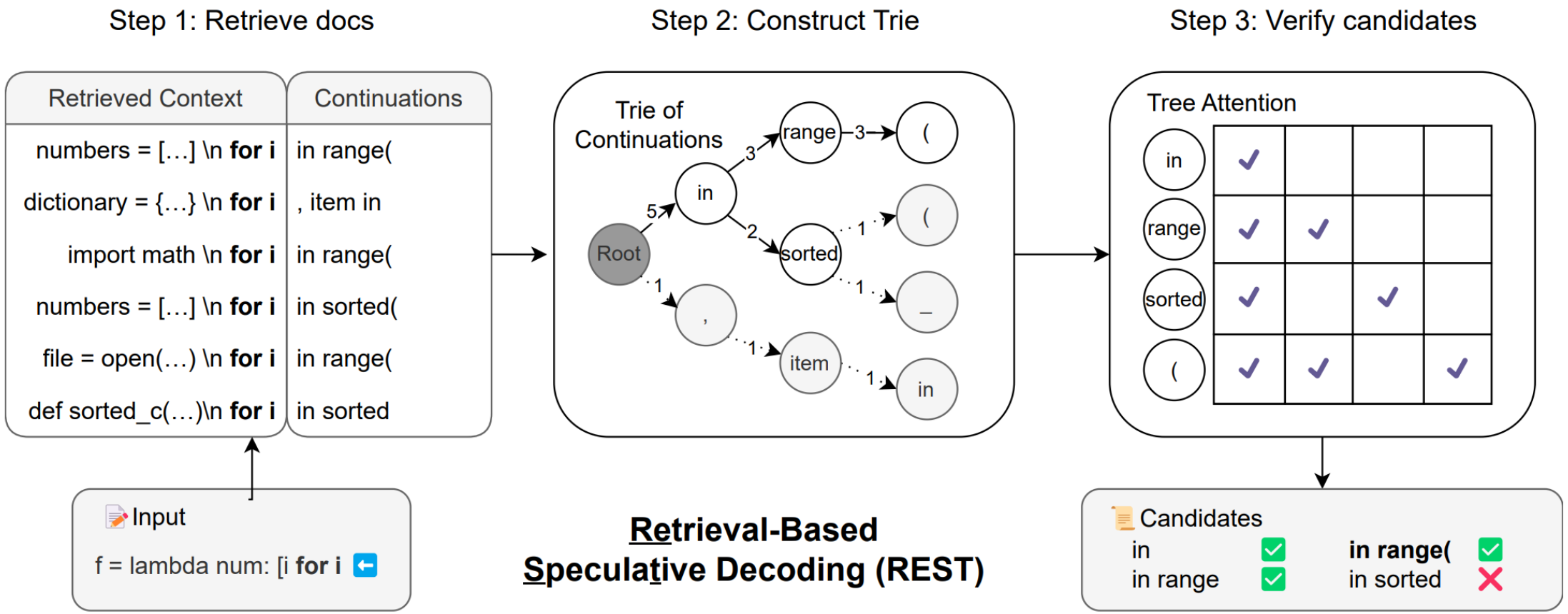


$$p(y|x) = \lambda\, p_{\text{kNN}}(y|x) + (1-\lambda)\, p_{\text{LM}}(y|x)$$

Khandelwal, Urvashi, et al. "Generalization through Memorization: Nearest Neighbor Language Models." International Conference on Learning Representations. 2019.

RA-LFM Learning: Training-free

- **REST**



He, Zhenyu, et al. "REST: Retrieval-Based Speculative Decoding." NAACL. 2024.

RA-LFM Learning: Training-free

- ✓ Work with off-the-shelf models
- x All components are fixed and not trained
- x Might not achieve optimal learning result of the whole model

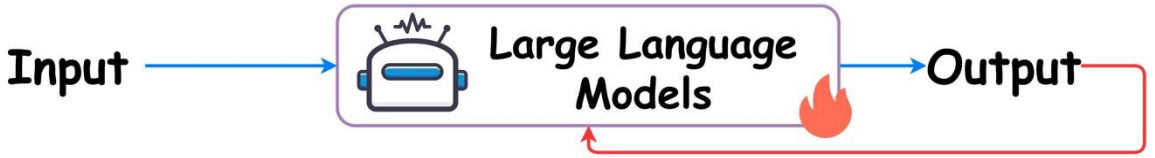
Part 3: RA-LFM Learning

Website of this tutorial

- Training-free Methods
- **Training-based Methods**
  - **Independent Learning**
  - Sequential Learning
  - Joint Learning

RA-LFM Learning: Independent Training

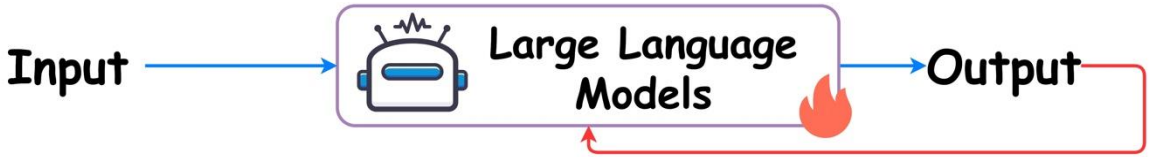
- **Retrieval models** and **language models** are trained **independently**.
  - **Independent training of large language models**.



  - **Independent training of Retriever**.

RA-LFM Learning: Independent Training

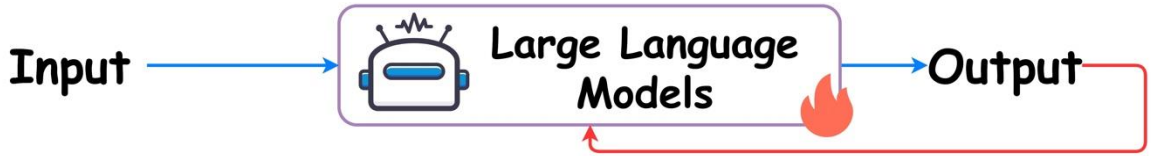
- **Retrieval models** and **language models** are trained **independently**.
  - **Independent training of large language models.**



  - **Independent training of Retriever.**

RA-LFM Learning: Independent Training

- **Independent training of large language models.**



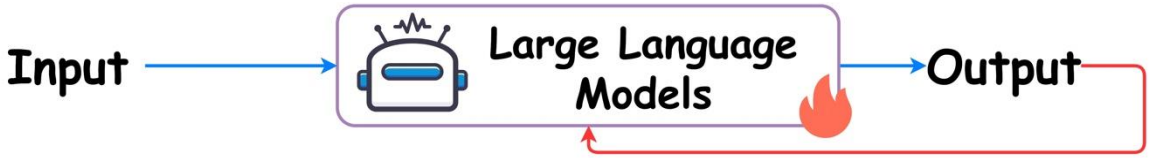
Back-Propagate

$$Minimize\ -logP_{LM}(y|x)$$

OpenAI Meta Google AI T5 .....

RA-LFM Learning: Independent Training

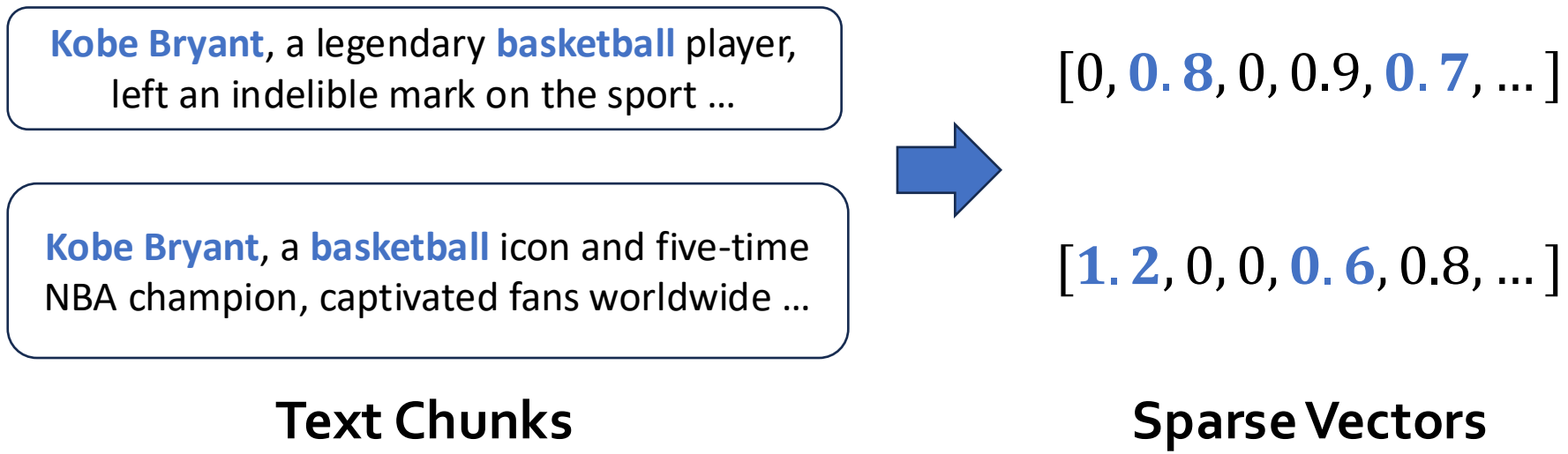
- **Retrieval models** and **language models** are trained **independently**.
  - **Independent training of large language models.**



- **Independent training of Retriever.**

RA-LFM Learning: Independent Training

- **Sparse retrieval models: TF-IDF / BM25**

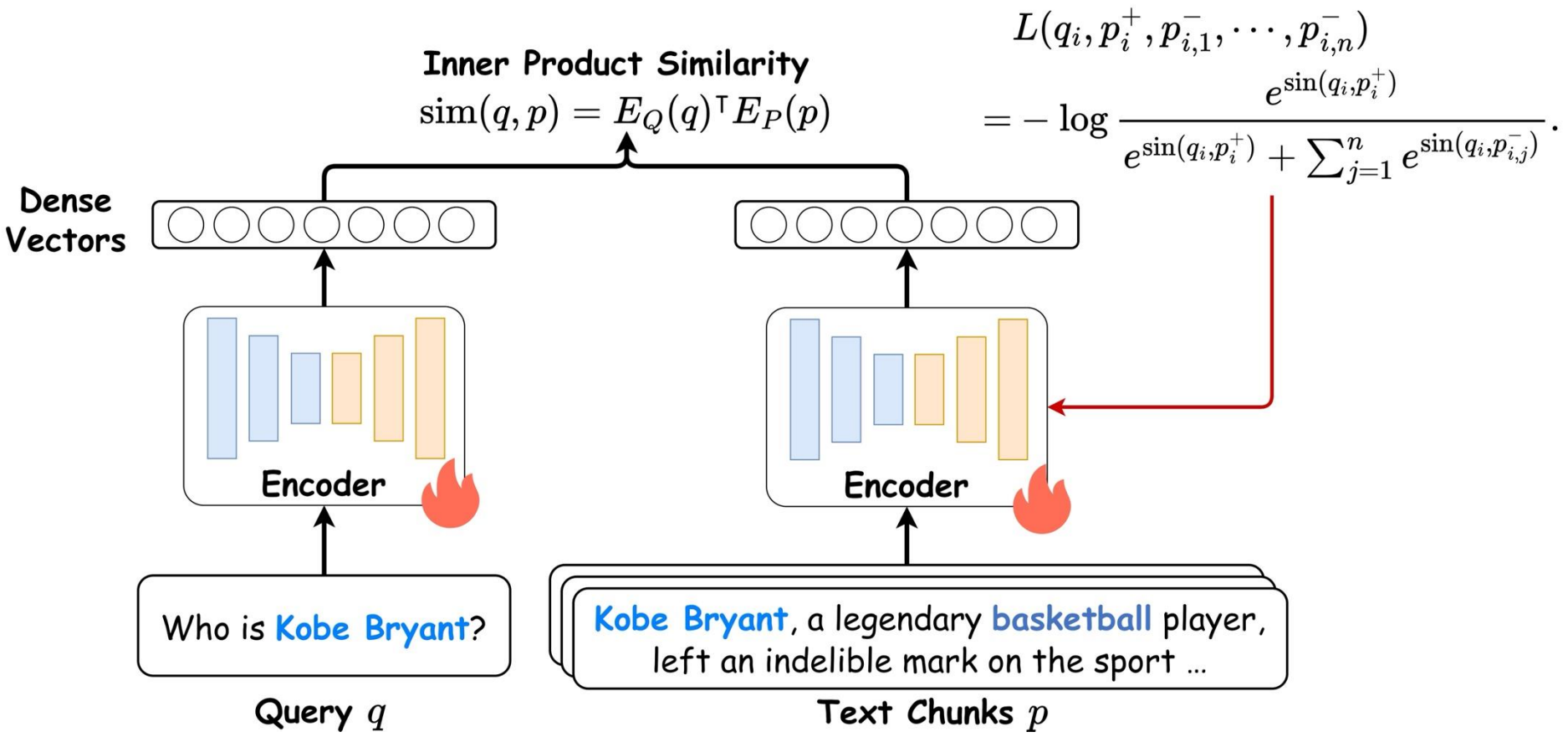


No training is Needed!

Ramos, Juan. "Using TF-IDF to determine word relevance in document queries." Proceedings of the first instructional conference on machine learning. 2003.
Robertson, Stephen, and Hugo Zaragoza. "The probabilistic relevance framework: BM25 and beyond." Foundations and Trends® in Information Retrieval. 2009.

RA-LFM Learning: Independent Training

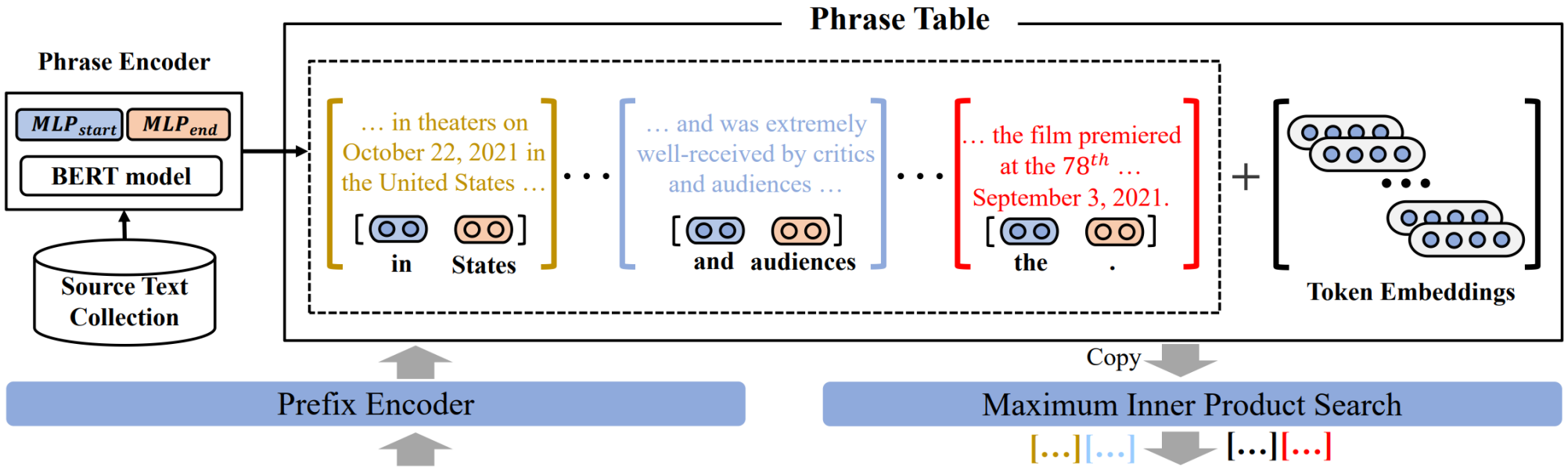
- **Dense retrieval models: DPR**



Karpukhin, Vladimir, et al. "Dense passage retrieval for open-domain question answering." 2020 Conference on Empirical Methods in Natural Language Processing, 2020.

RA-LFM Learning: Independent Training

- **Dense retrieval models: CoG**



$$\mathcal{H}_{i+1} = \text{PrefixEncoder}(x_i, \mathcal{H}_i).$$

$$\mathcal{D}_{\text{start}} = \text{MLP}_{\text{start}}(\mathcal{D}), \mathcal{D}_{\text{end}} = \text{MLP}_{\text{end}}(\mathcal{D}).$$

$$\text{PhraseEncoder}(s, e, D) = [\mathcal{D}_{\text{start}}[s]; \mathcal{D}_{\text{end}}[e]] \in \mathbb{R}^d$$

Tian Lan, et al. "Copy is All You Need." In The Eleventh International Conference on Learning Representations, 2022.

RA-LFM Learning: Independent Training

- **Model Training:**

$$\mathcal{L}_p = -\frac{1}{n}\sum_{k=1}^{n}\log\frac{\exp(q_k \cdot p_k)}{\sum_{p\in\mathcal{P}_k}\exp(q_k \cdot p_p) + \sum_{w\in V}\exp(q_k \cdot v_w)}$$

$$\mathcal{L}_t = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{\exp(q_i, v_{D_i})}{\sum_{w\in V}\exp(q_i, v_w)}$$

Tian Lan, et al. “Copy is All You Need.” In The Eleventh International Conference on Learning Representations, 2022.

RA-LFM Learning: Independent Training

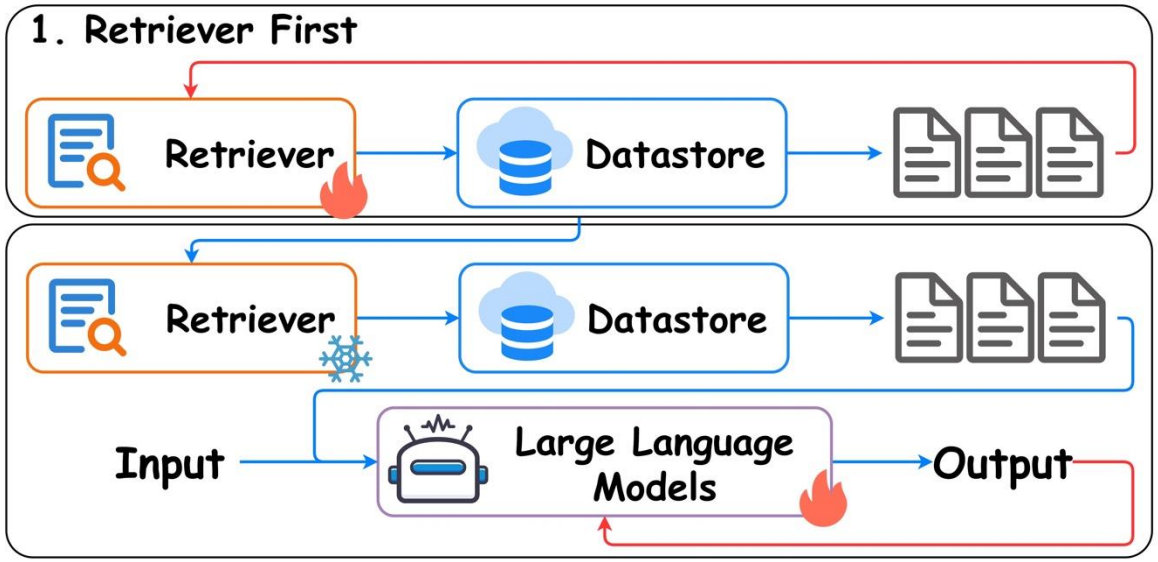
- ✓ Work with off-the-shelf models, flexible
- ✓ Each part can be improved independently
- x Lack of integrity between Retrieval and Generation
- x Retrieval models are not optimized specified for the tasks/ domains/ generators

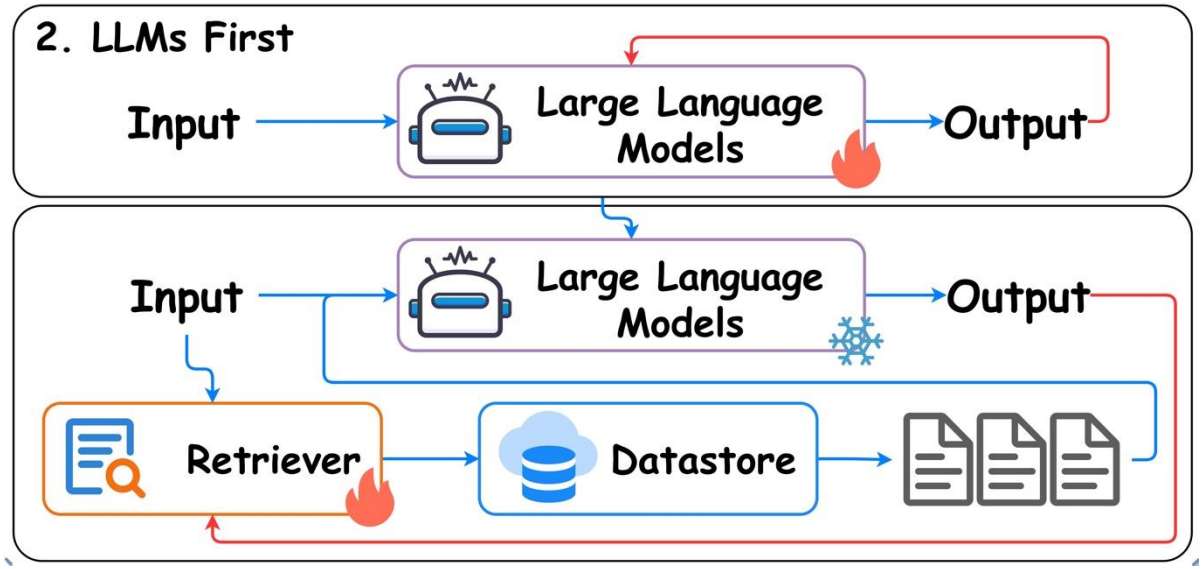
Part 3: RA-LFM Learning

- Training-free Methods
- **Training-based Methods**
  - Independent Learning
  - **Sequential Learning**
  - Joint Learning

RA-LFM Learning: Sequential Training

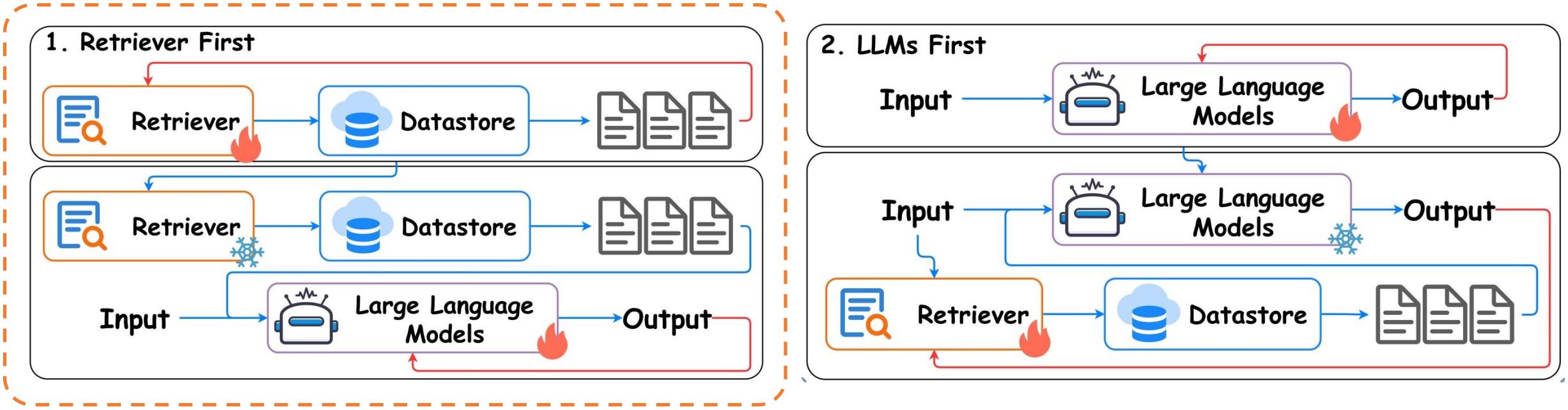
- **One component** is first trained independently and then fixed.
- **The other component** is trained with an objective that depends on **the first one.**





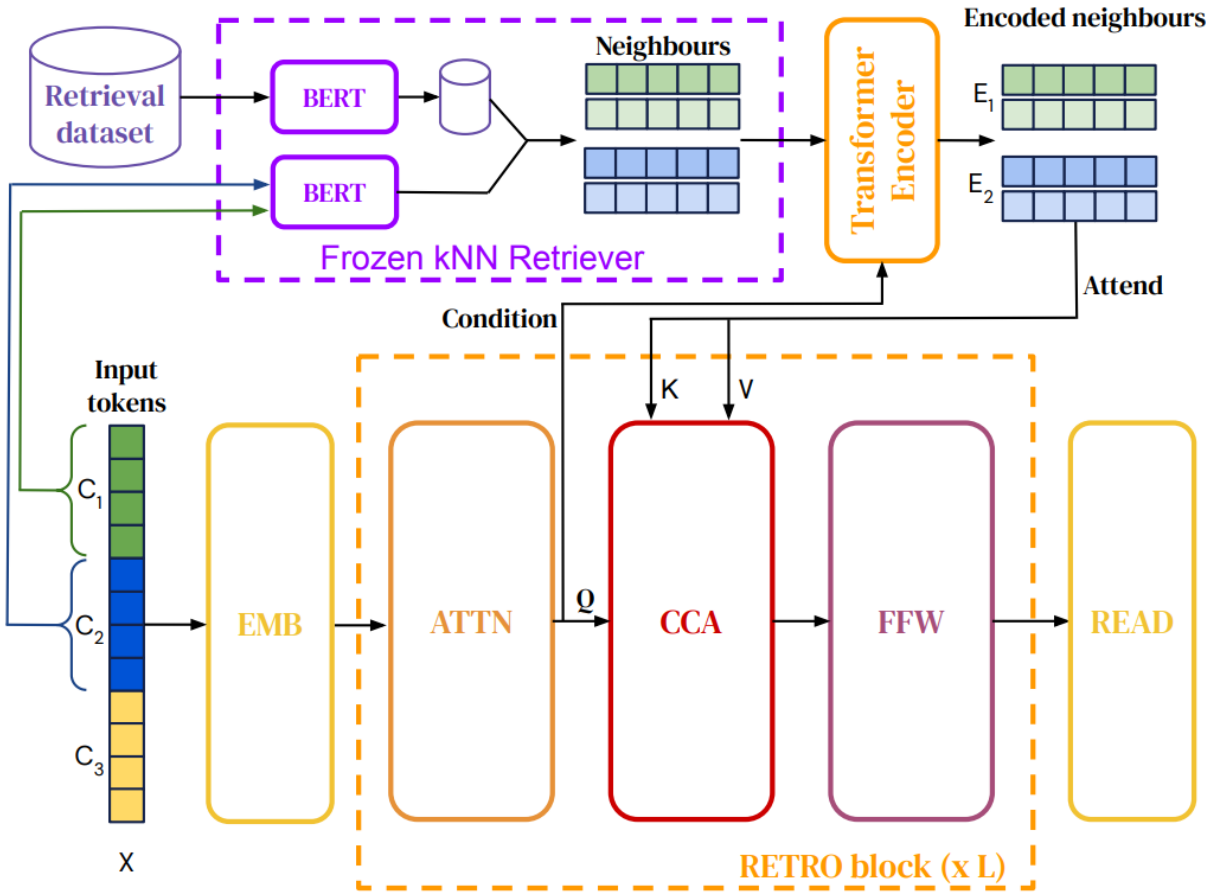
RA-LFM Learning: Sequential Training

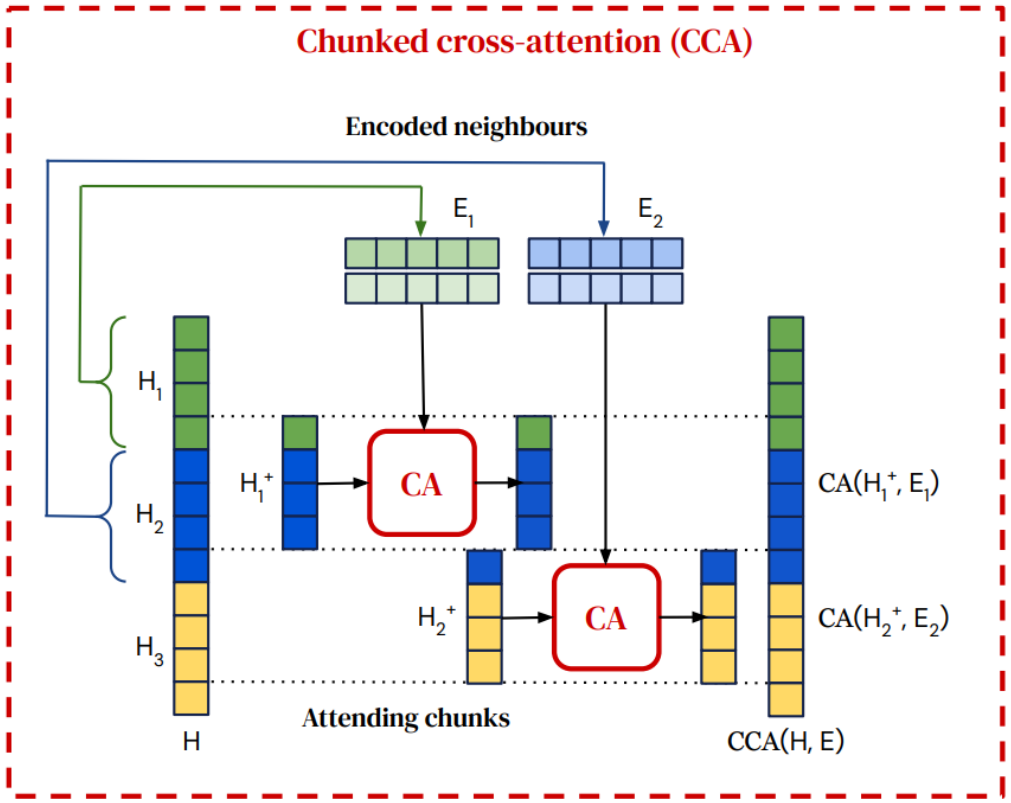
- **Retrieval models** is first trained independently and then fixed.
- **Language models** are trained with an objective that depends on **the Retrieval.**



RA-LFM Learning: Sequential Training

- **RETRO**

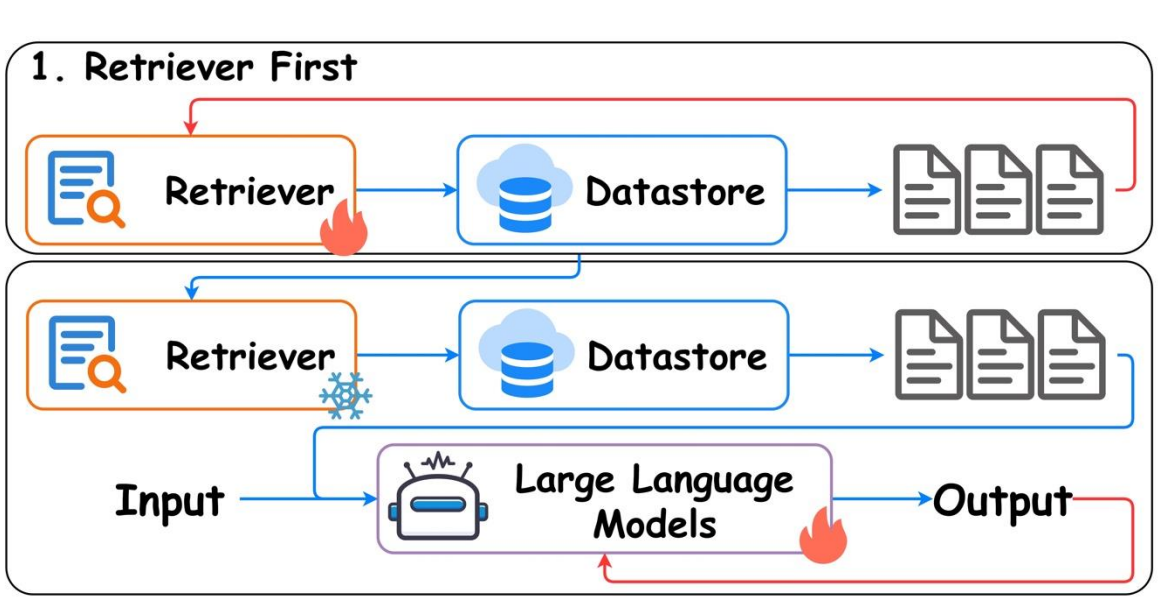


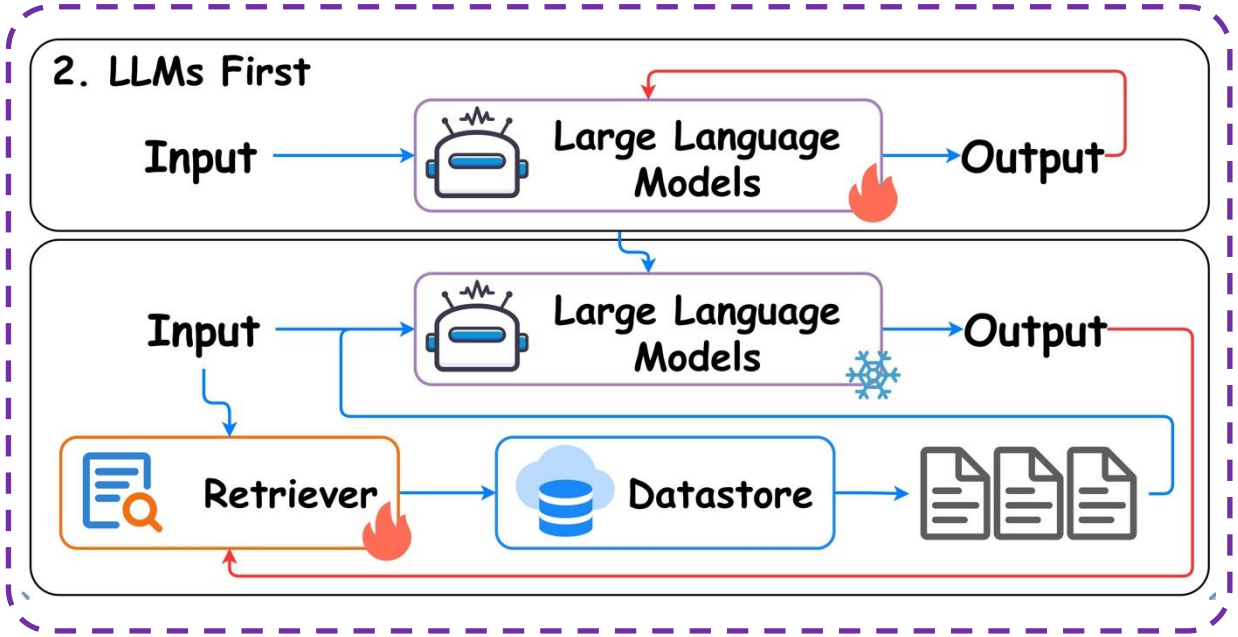


Borgeaud et al. 2022. Improving Language Models by Retrieving from Trillions of Tokens

RA-LFM Learning: Sequential Training

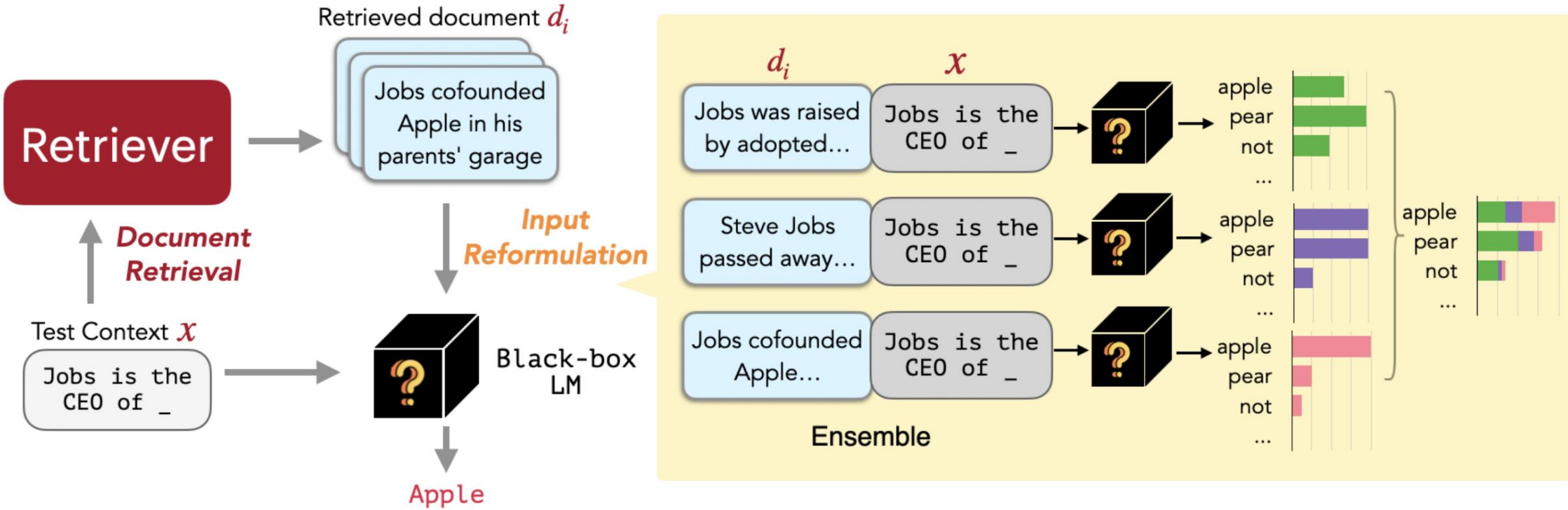
- **Language models** are first trained independently and then fixed.
- **Retrieval models** are trained with supervisions from **language models.**





RA-LFM Learning: Sequential Training

- **REPLUG (Retrieve and Plug)**

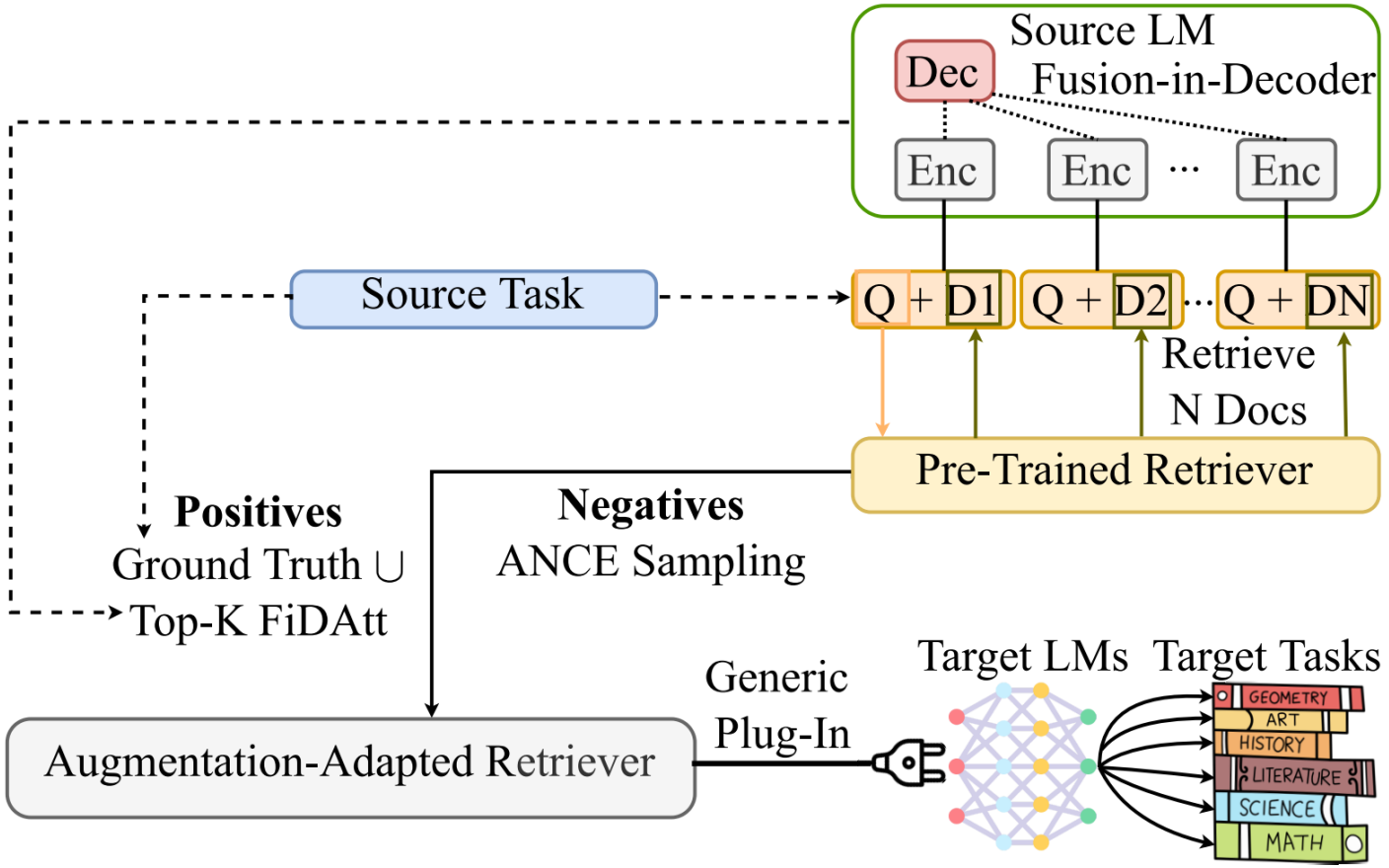


$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} KL\Big(P_R(d \mid x) \parallel Q_{\text{LM}}(d \mid x, y)\Big) \qquad P_R(d \mid x) = \frac{e^{s(d,x)/\gamma}}{\sum_{d \in \mathcal{D}'} e^{s(d,x)/\gamma}} \qquad Q(d \mid x, y) = \frac{e^{P_{LM}(y|d,x)/\beta}}{\sum_{d \in \mathcal{D}'} e^{P_{LM}(y|d,x)/\beta}}$$

Shi, Weijia, et al. "REPLUG: Retrieval-Augmented Black-Box Language Models." NAACL. 2024.

RA-LFM Learning: Sequential Training

- **AAR (Augmentation-Adapted Retriever)**



$$\mathcal{L} = \sum_{q} \sum_{d^+ \in D^{a+}} \sum_{d^- \in D^-} l(f(\boldsymbol{q}, \boldsymbol{d}^+), f(\boldsymbol{q}, \boldsymbol{d}^-)),$$

Yu, Zichun, et al. "Augmentation-Adapted Retriever Improves Generalization of Language Models as Generic Plug-In." ACL. 2023.

RA-LFM Learning: Sequential Training

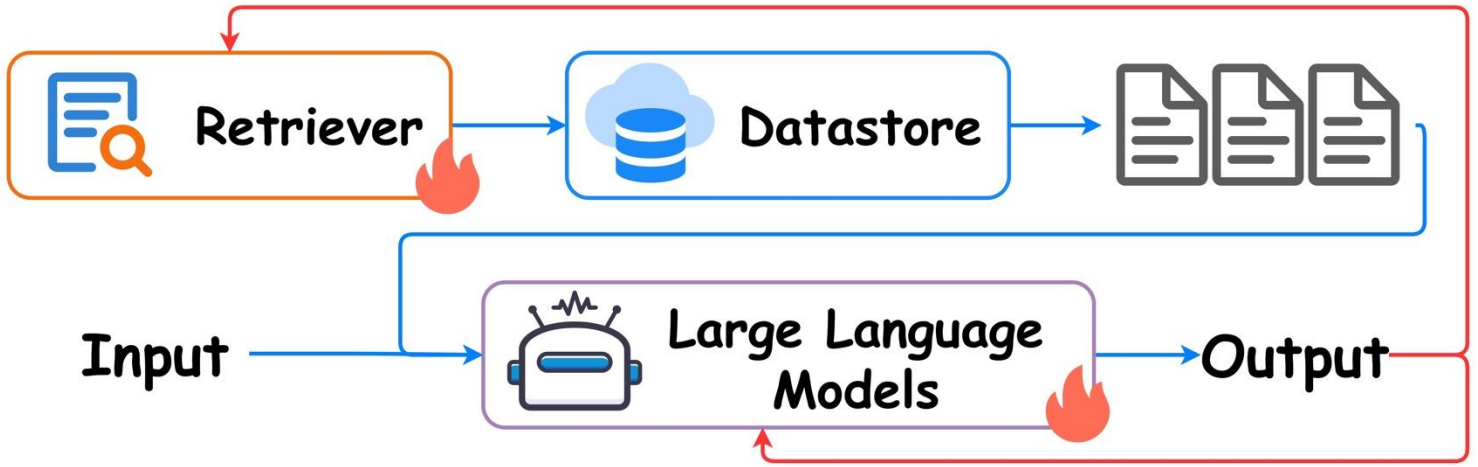
- ✓ Work with off-the-shelf models
- ✓ Generators can be trained effectively based on the retrieved results
- ✓ Retrievers can be trained to provide useful information to help the generators
- x One component is still fixed and not trained
- x Might not achieve optimal learning result of the whole model

Part 3: RA-LFM Learning

- Training-free Methods
- **Training-based Methods**
  - Independent Learning
  - Sequential Learning
  - **Joint Learning**

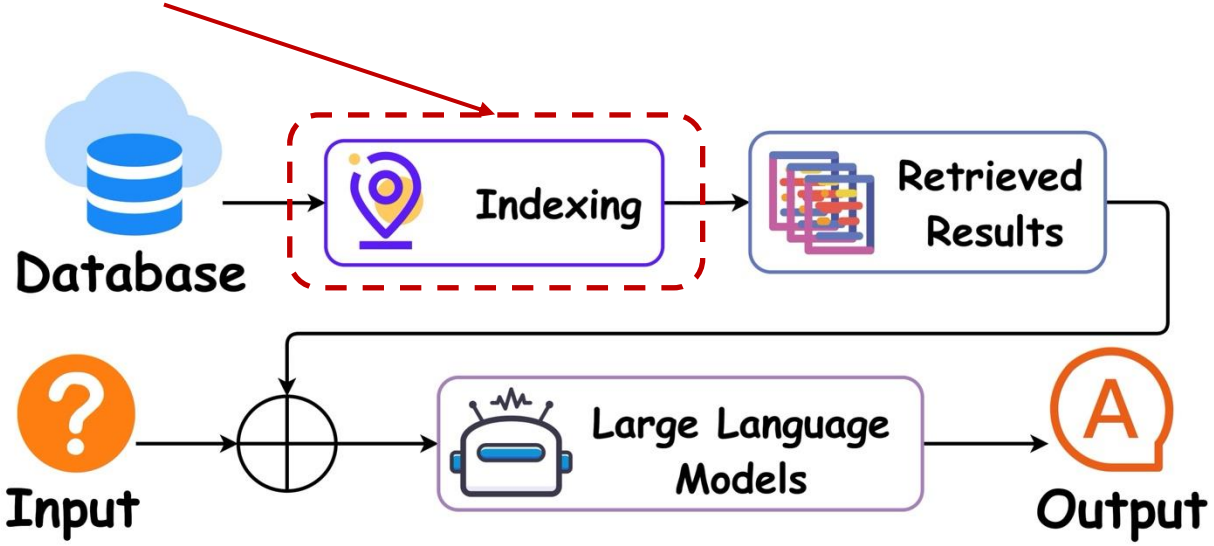
RA-LFM Learning: Joint Training

- **Retrieval models** is and **language models** are trained jointly.



RA-LFM Learning: Joint Training

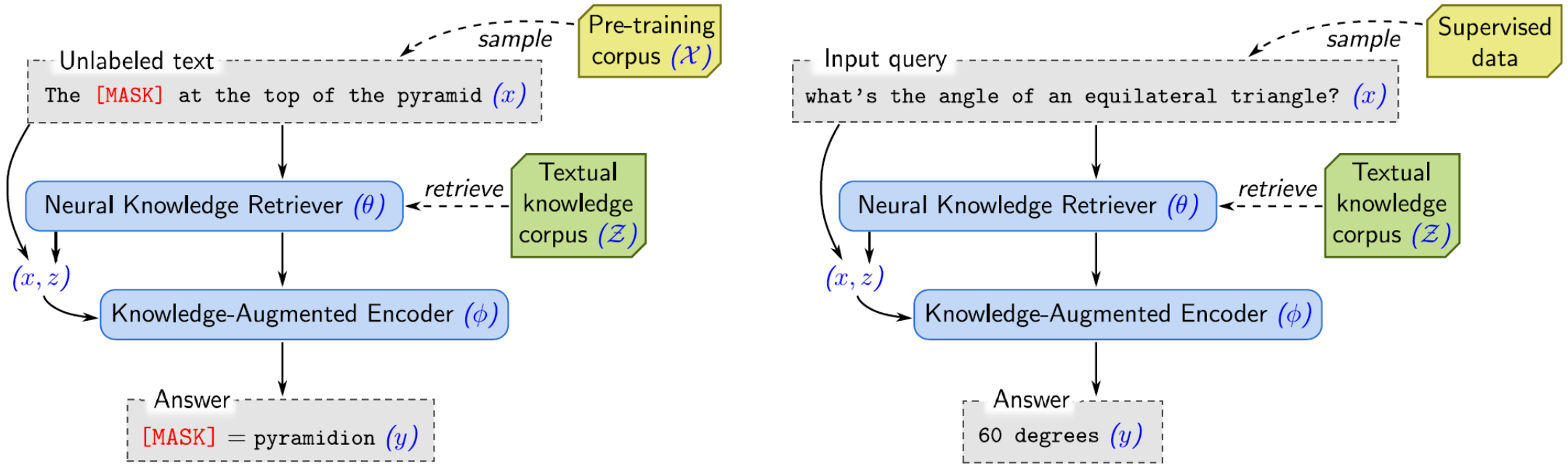
- **Retrieval Index Updating, which could be very expensive!**



- **Solutions:**
  - **Asynchronous index updating**
  - **In-batch approximation**

RA-LFM Learning: Joint Training

- **REALM**

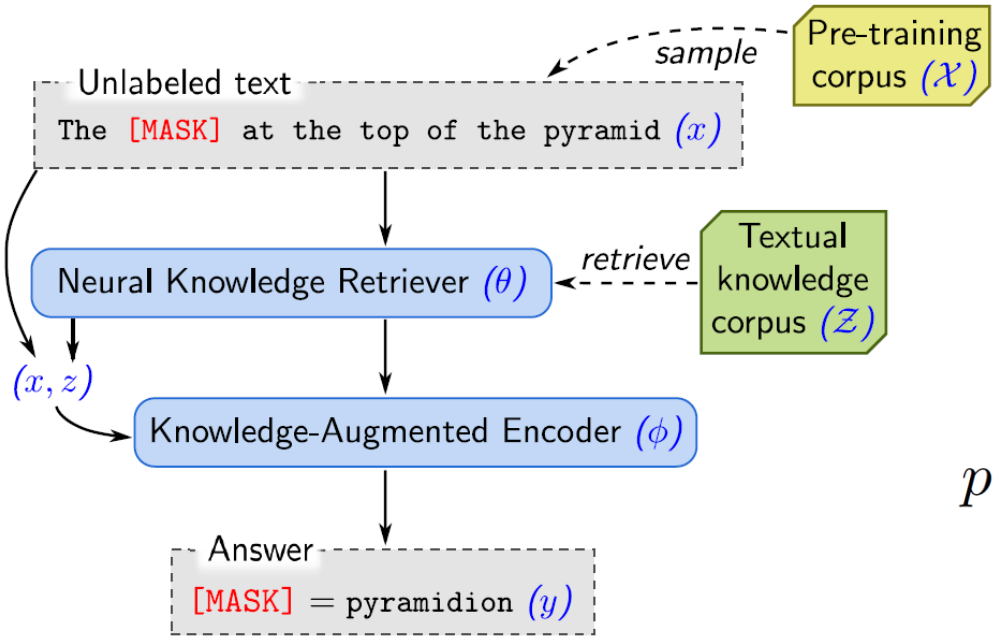


Objective function: $$p(y \mid x) = \sum_{z \in \mathcal{Z}} p(y \mid z, x)\, p(z \mid x).$$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

RA-LFM Learning: Joint Training

- **REALM**



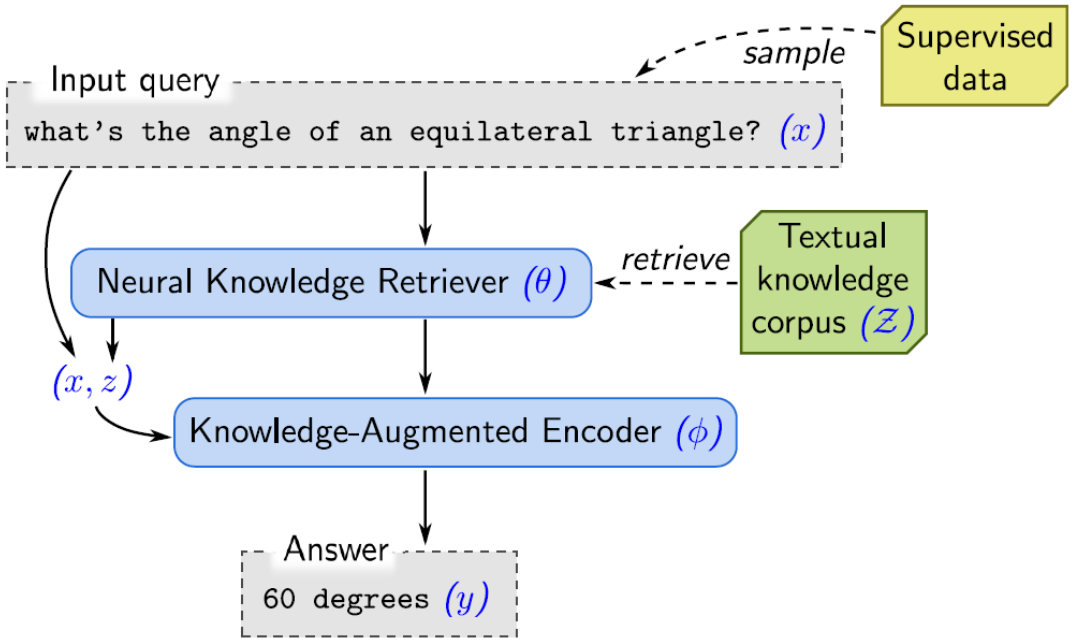
$$p(y \mid z, x) = \prod_{j=1}^{J_x} p(y_j \mid z, x)$$

$$p(y_j \mid z, x) \propto \exp\left(w_j^\top \texttt{BERT}_{\texttt{MASK}(j)}(\texttt{join}_{\texttt{BERT}}(x, z_{\text{body}}))\right)$$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

RA-LFM Learning: Joint Training

- **REALM**



$$p(y \mid z, x) \propto \sum_{s \in S(z,y)} \exp\left(\texttt{MLP}\left(\left[h_{\texttt{START(s)}}; h_{\texttt{END(s)}}\right]\right)\right)$$

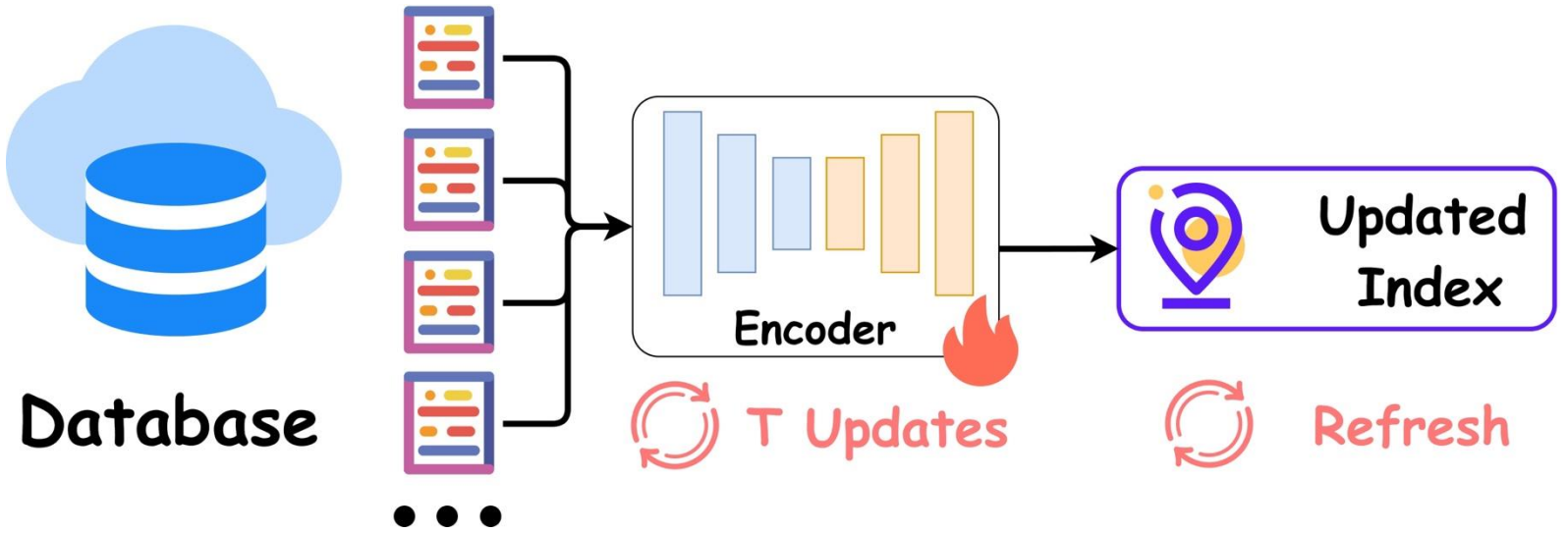
$$h_{\texttt{START(s)}} = \texttt{BERT}_{\texttt{START(s)}}(\texttt{join}_{\texttt{BERT}}(x, z_{\text{body}})),$$

$$h_{\texttt{END(s)}} = \texttt{BERT}_{\texttt{END(s)}}(\texttt{join}_{\texttt{BERT}}(x, z_{\text{body}})),$$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

RA-LFM Learning: Joint Training

- **REALM – Asynchronous Index Update**

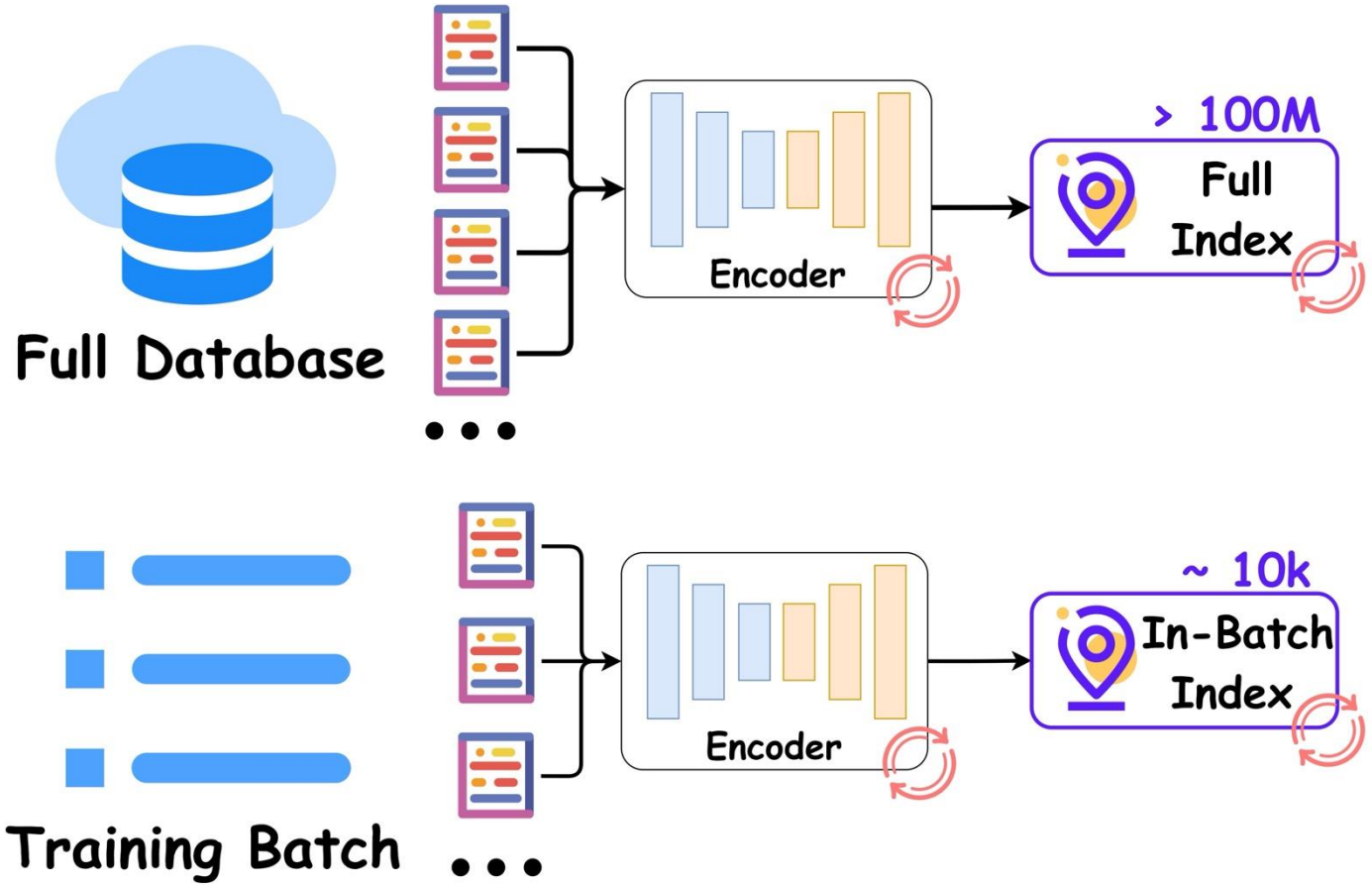


$$f(x, z) = \text{Embed}_{\text{input}}(x)^\top \text{Embed}_{\text{doc}}(z)$$

Guu et al. 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

RA-LFM Learning : Joint Training

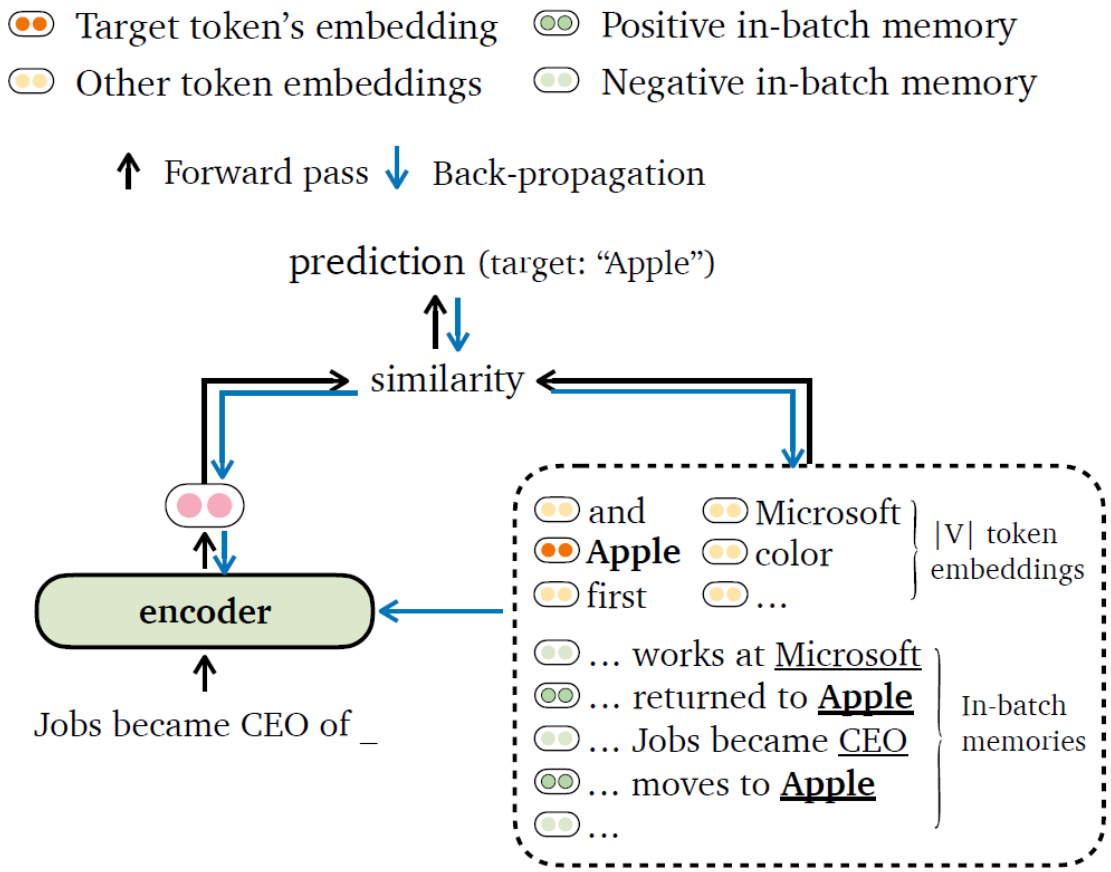
- **TRIME – In-Batch Approximation**



Zhong et al., 2022. “Training Language Models with Memory Augmentation”

RA-LFM Learning : Joint Training

- **TRIME**



Local Memory: $\mathcal{M}_{\text{local}}(c_t) = \{(c_j, x_j)\}_{1 \le j \le t-1}$.

Long-term Memory:

$$\mathcal{M}_{\text{long}}(c_t^{(i)}) = \{(c_j^{(k)}, x_j^{(k)})\}_{1 \le k < i, 1 \le j}$$

External Memory: $\mathcal{M}_{\text{ext}} = \{(c_j, x_j) \in \mathcal{D}\}$.

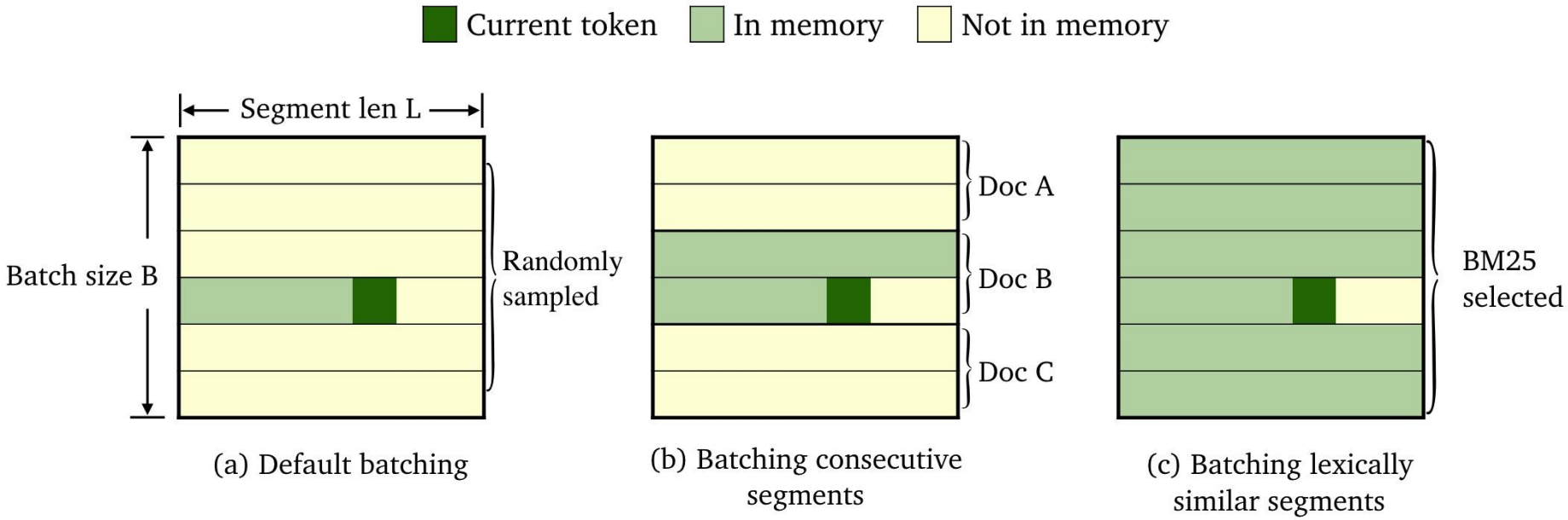
Training Objective:

$$P(w \mid c) \propto \exp(E_w^\top f_\theta(c)) + \sum_{(c_j, x_j) \in \mathcal{M}_{\text{train}}: x_j = w} \exp(\text{sim}(g_\theta(c), g_\theta(c_j))).$$

Zhong et al., 2022. "Training Language Models with Memory Augmentation"

RA-LFM Learning : Joint Training

- **TRIME Data Batching Strategy**



Use BM25 scores to find similar text chunks to provide more training signals

Zhong et al., 2022. "Training Language Models with Memory Augmentation"

Tutorial Outline

- **Part 1: Introduction** of Retrieval Augmented Large Foundation Models (RA-LFMs) (Dr. Yujuan Ding)
- **Part 2: Architecture** of RA-LFMs and **Main Modules** (Xu Yuan)
- **Part 3: Learning Approach of RA-LFMs (Chengliang Liu)**
- **Part 4: Agentic RAG (Chengliang Liu)**
- **Part 5: Applications** of RA-LFMs (Chun-Hin Chan)
- **Part 6: Challenges and Future Directions** of RA-LFMs (Dr. Yujuan Ding)
- **Part 7: Q&A**

Website of this tutorial
Check out the slides and more information!

Part 4: Agentic RAG

Presenter
Chengliang Liu
HK PolyU

- **Motivation, Definition & Key Features**
- Representative Methods
  - Prompting
  - Single-Agent
  - Multi-Agent
  - RL-Driven

Motivation, Definition & Key Features

Static pipeline: Query → Retrieve → Reason → Answer (fixed, one-shot)

Three limitations of this paradigm:

1. **Retrieval Adequacy:** pre-retrieved knowledge may not cover needs that emerge during reasoning
2. **Reasoning Depth:** noisy or wrong retrieval interferes with multi-step inference
3. **Adaptability:** no mechanism to adjust retrieval strategy mid-reasoning

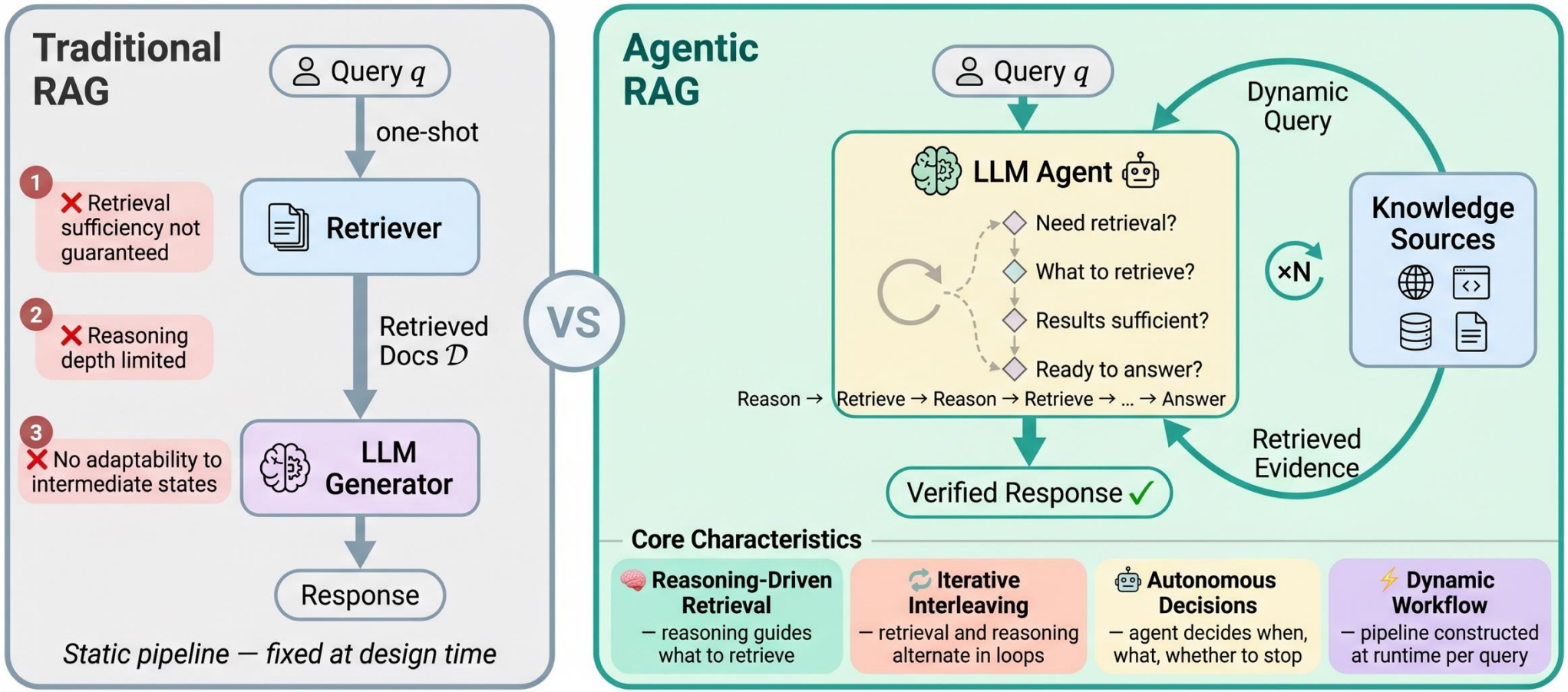
Motivation, Definition & Key Features

Agentic RAG: Query → [Agent loop: Reason ↔ Retrieve ↔ Evaluate] → Answer

1. **Reasoning-Driven Retrieval**
   Reasoning process triggers and guides retrieval direction
2. **Iterative Interleaving**
   Retrieval and reasoning alternate in multiple rounds (RAG ⇔ Reasoning closed loop)
3. **Autonomous Decision-Making**
   Agent judges at runtime: when to retrieve, what to retrieve, when to stop
4. **Dynamic Workflow**
   Execution path constructed per query, not pre-defined

Li et al. Towards Agentic RAG with Deep Reasoning: A Survey of RAG-Reasoning Systems in LLMs, EMNLP, 2025.

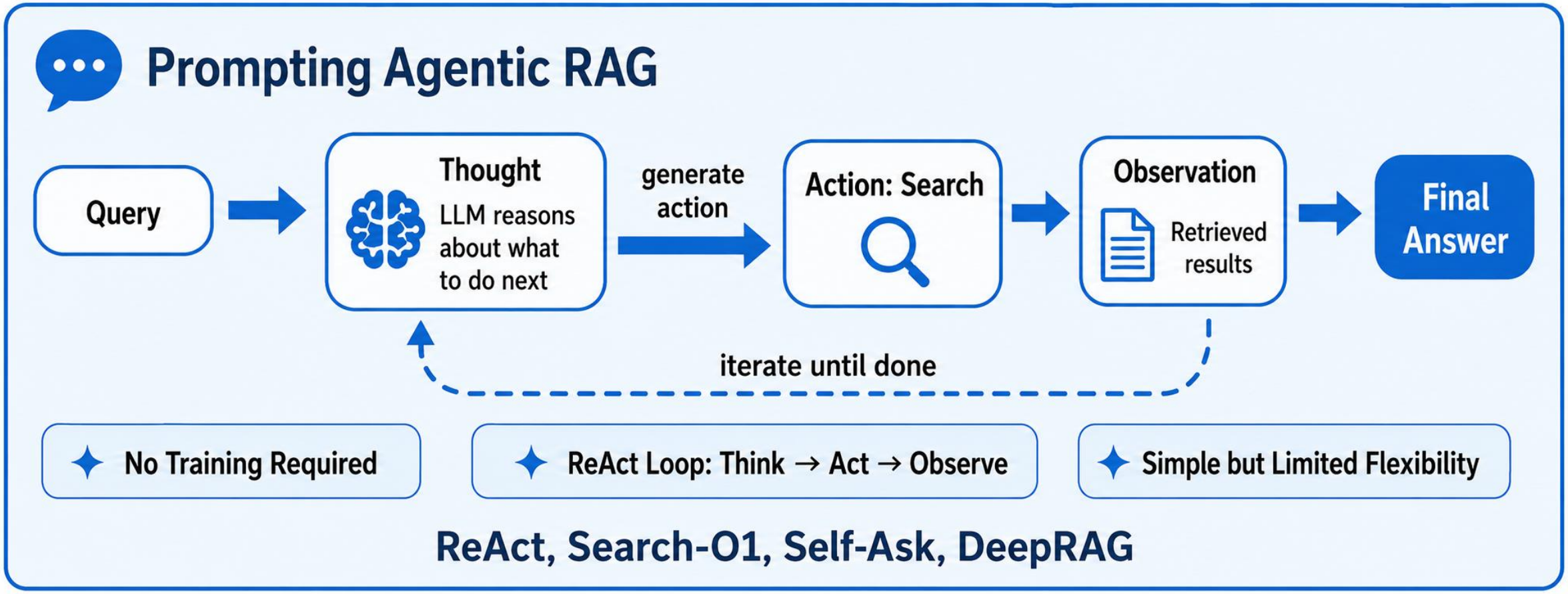
Motivation, Definition & Key Features



Representative Methods

Approach	How Agentic Behavior Arises	Key Characteristics
Prompting	Prompt design elicits reasoning-action loops	Zero training; flexible; prompt-dependent
Single-Agent SFT	Train model to internalize retrieval decisions	On-demand retrieval; self-reflection
Multi-Agent	Specialized agents collaborate on sub-tasks	Parallel; modular; heterogeneous sources
RL-Driven	Reward signals let model discover search strategies	Emergent behaviors; strong generalization

Prompt Chaining Agentic RAG



Prompt Chaining Agentic RAG

(1) Hotspot QA

Question: Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

(1a) Standard

Answer: iPod ✗

(1b) CoT (Reason Only)

Thought: Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

Answer: iPhone, iPad, iPod Touch ✗

(1c) Act-Only

Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control …

Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: …

Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Act 4: Finish[yes] ✗

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

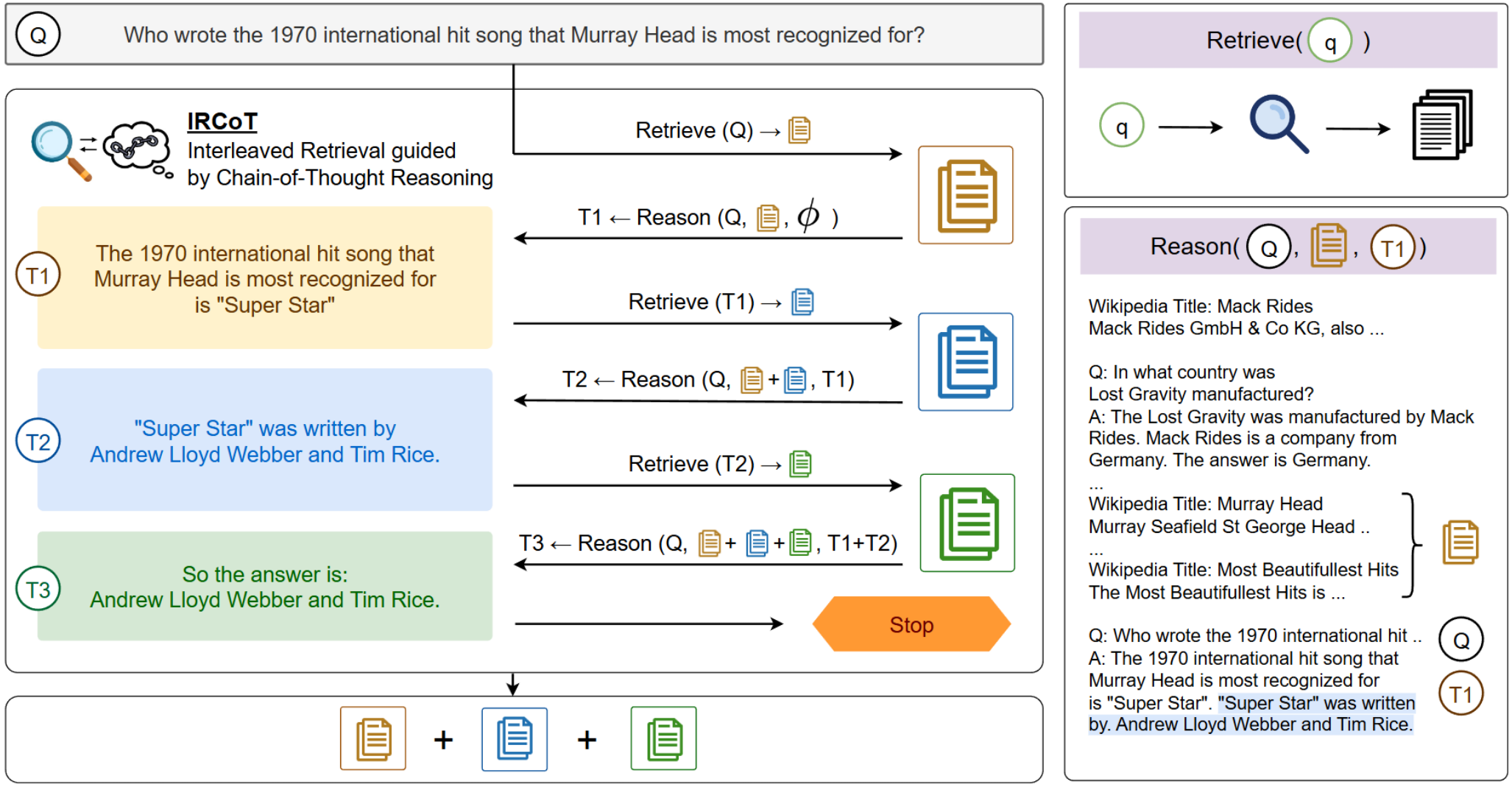
Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports',' Front Row (software) ', …]

Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys] ✓

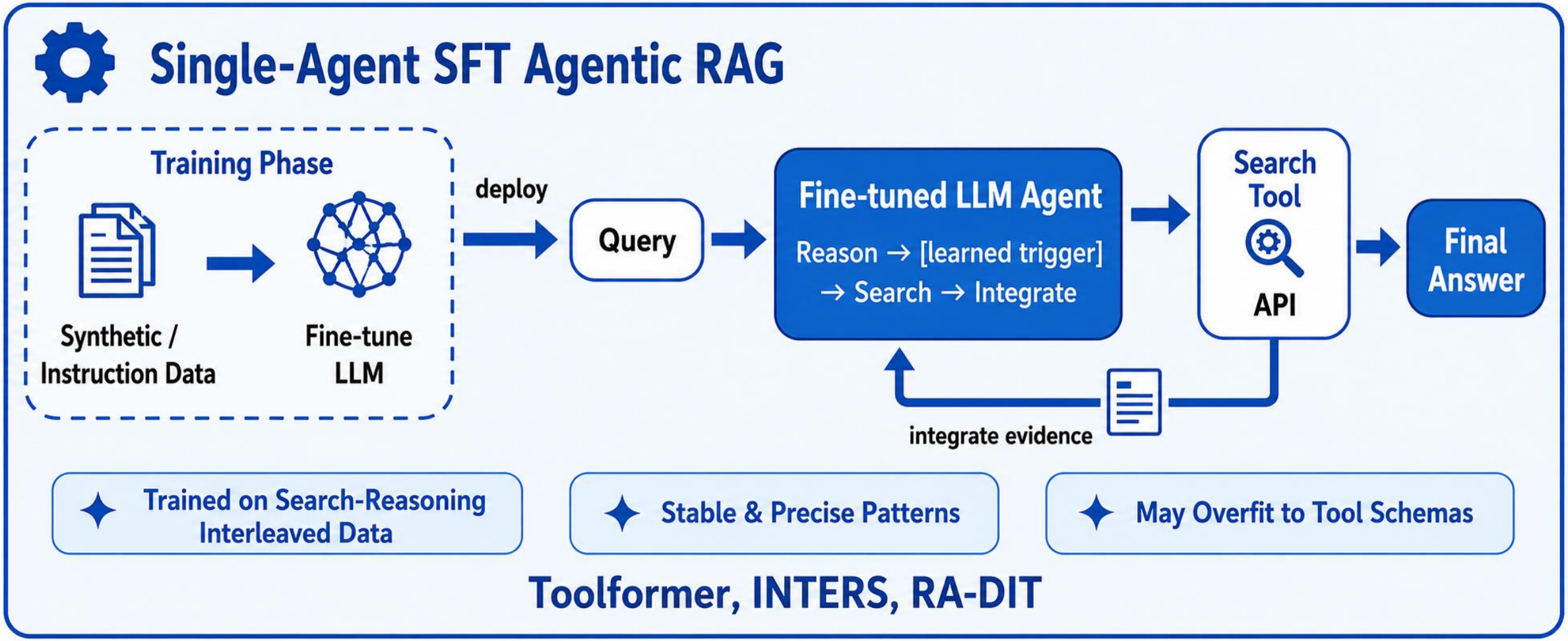
Yao et al. ReAct: Synergizing Reasoning and Acting in Language Models, ICLR, 2023.

Prompt Chaining Agentic RAG

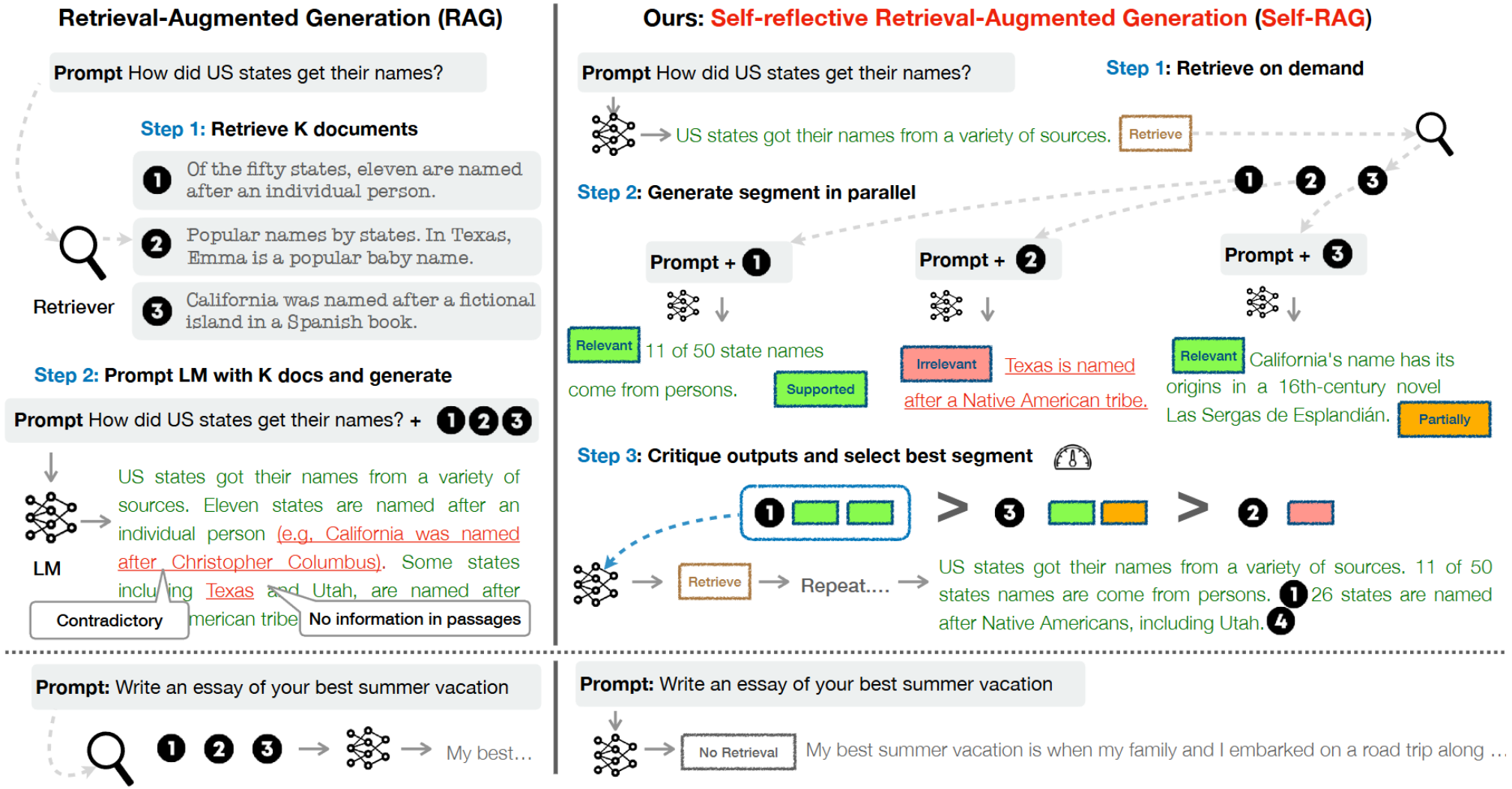


Trivedi et al. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions, ACL, 2023.

Single-Agent Agentic RAG

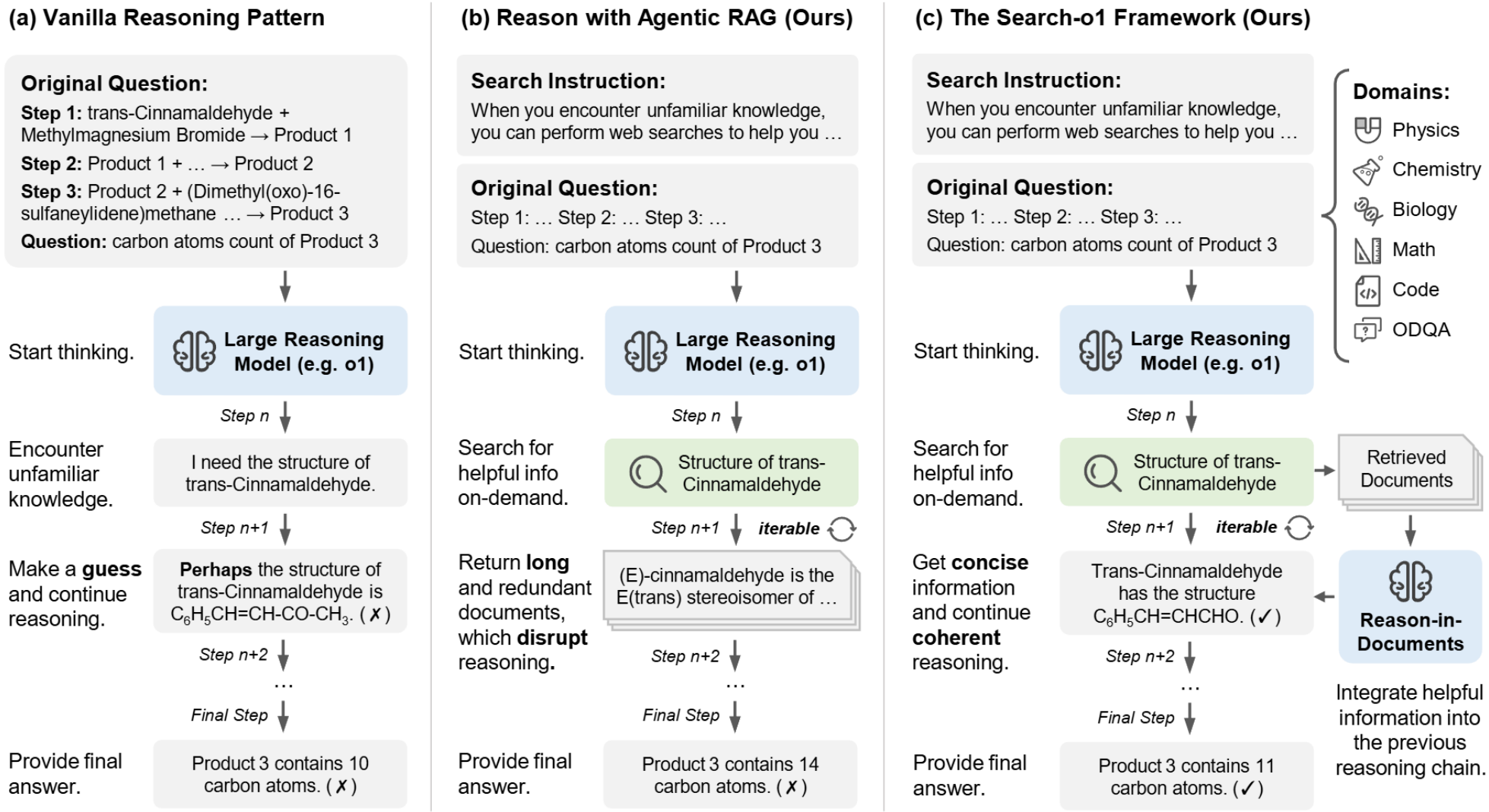


Single-Agent Agentic RAG



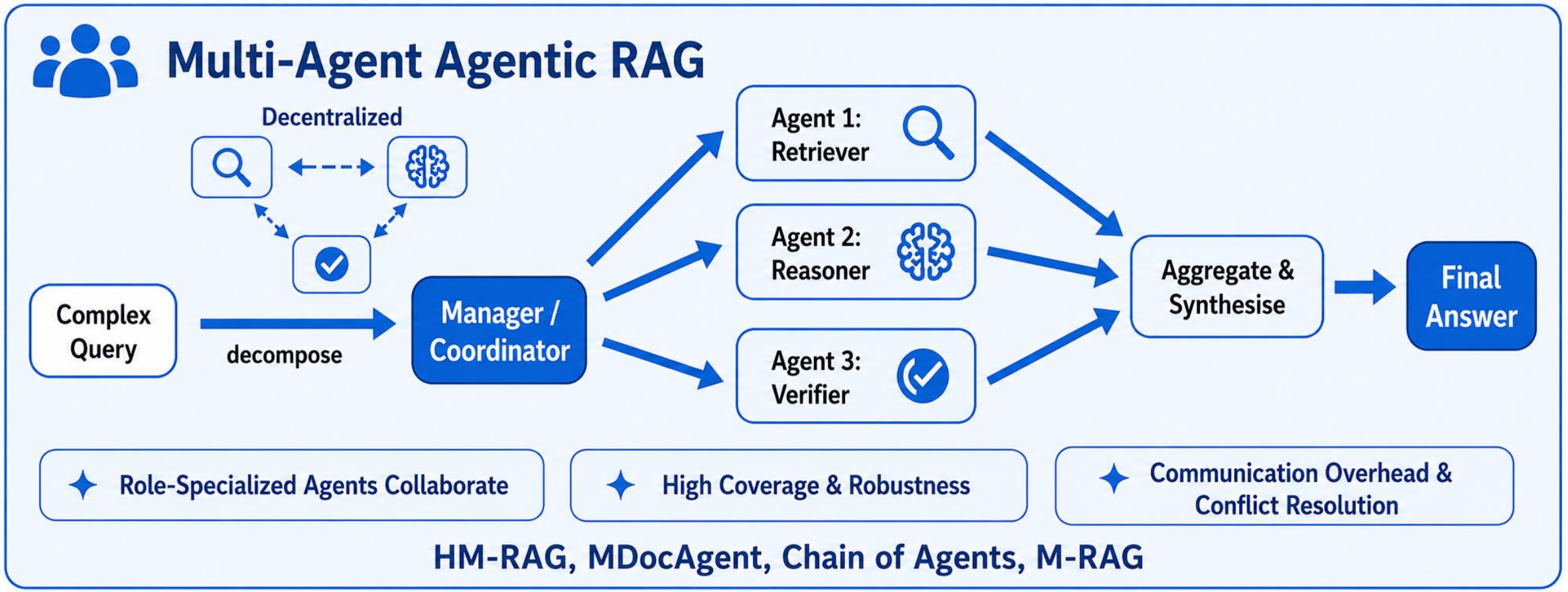
Asai et al. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection, ICLR, 2024.

Single-Agent Agentic RAG

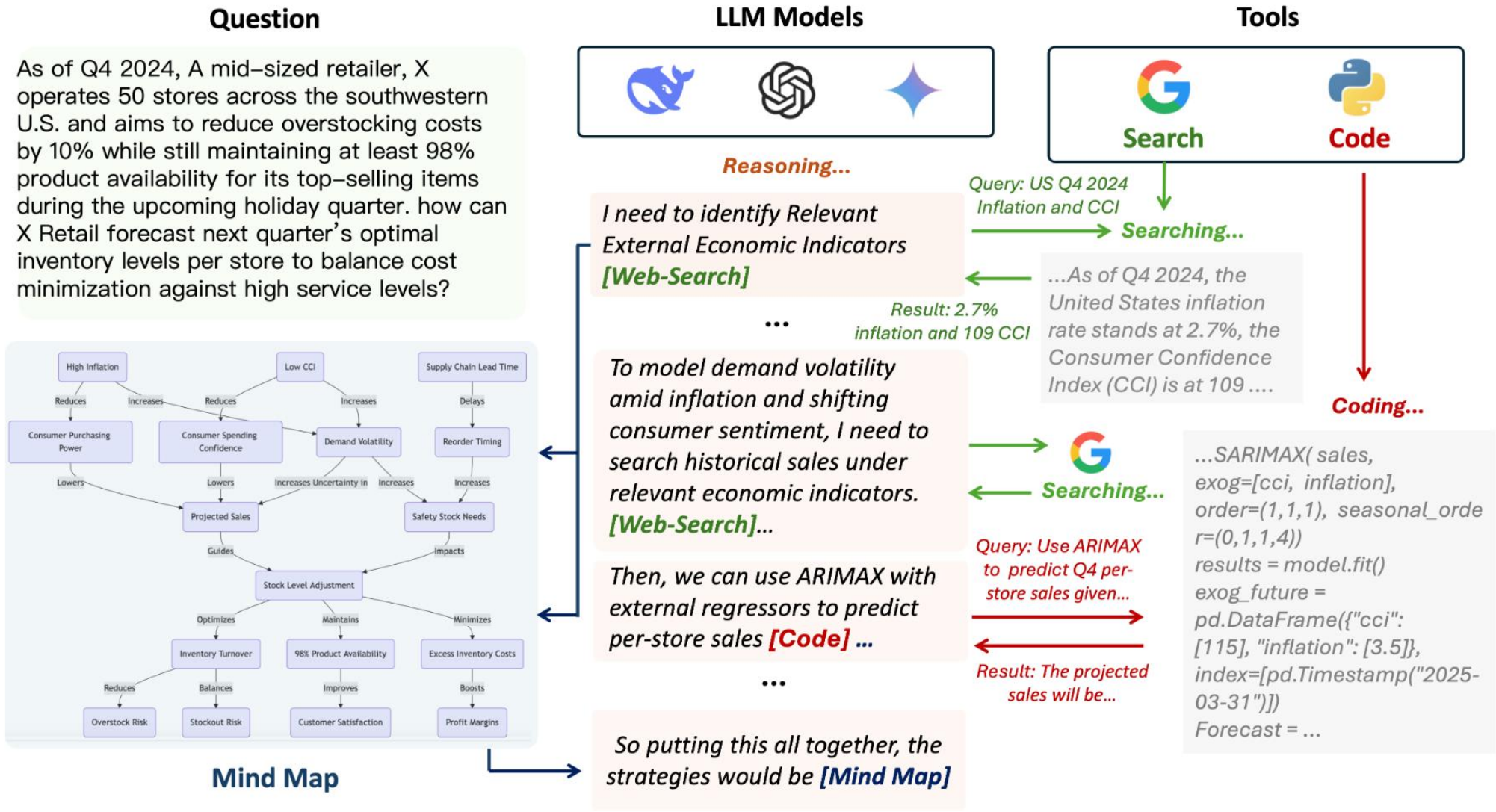


Li et al. Search-O1: Agentic Search-Enhanced Large Reasoning Models, arXiv, 2025.

Multi-Agent Agentic RAG

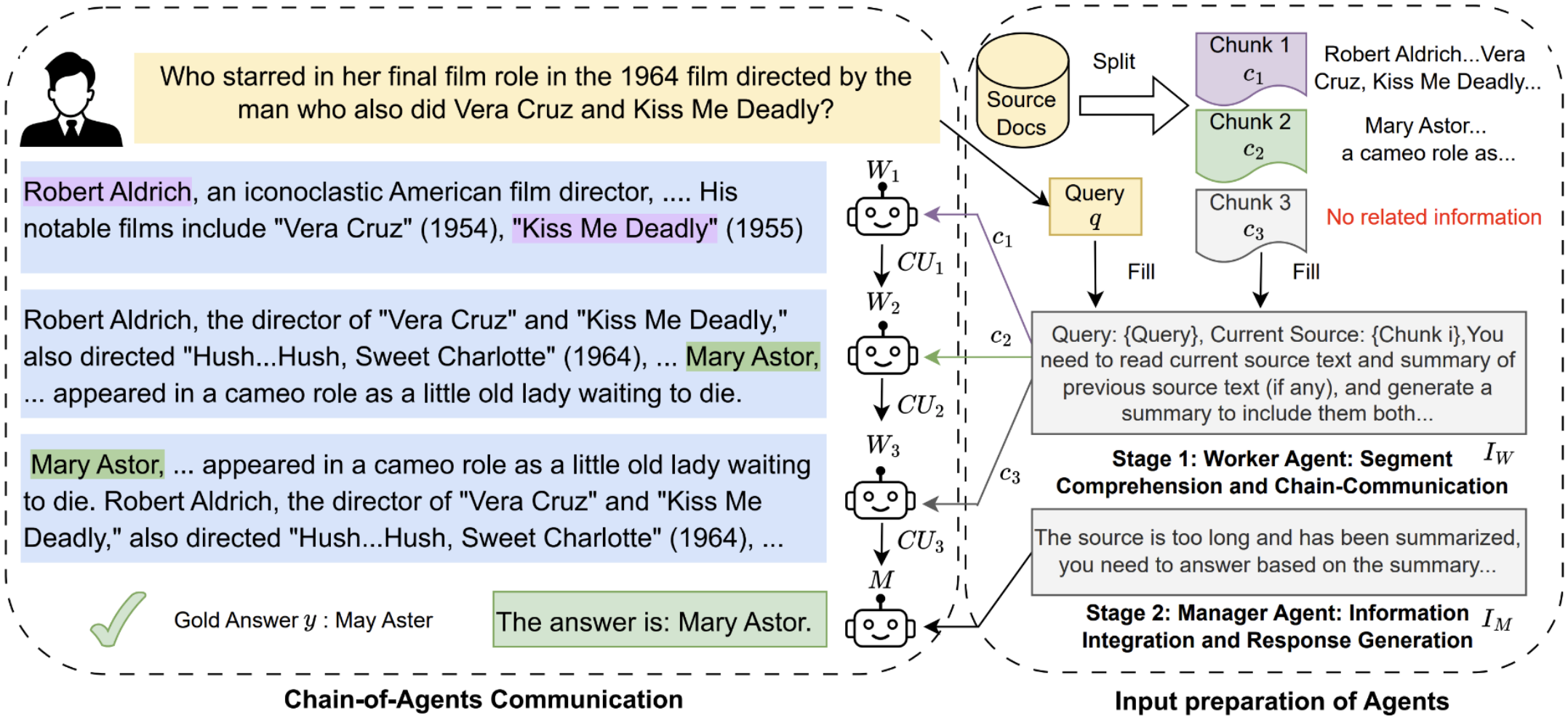


Multi-Agent Agentic RAG



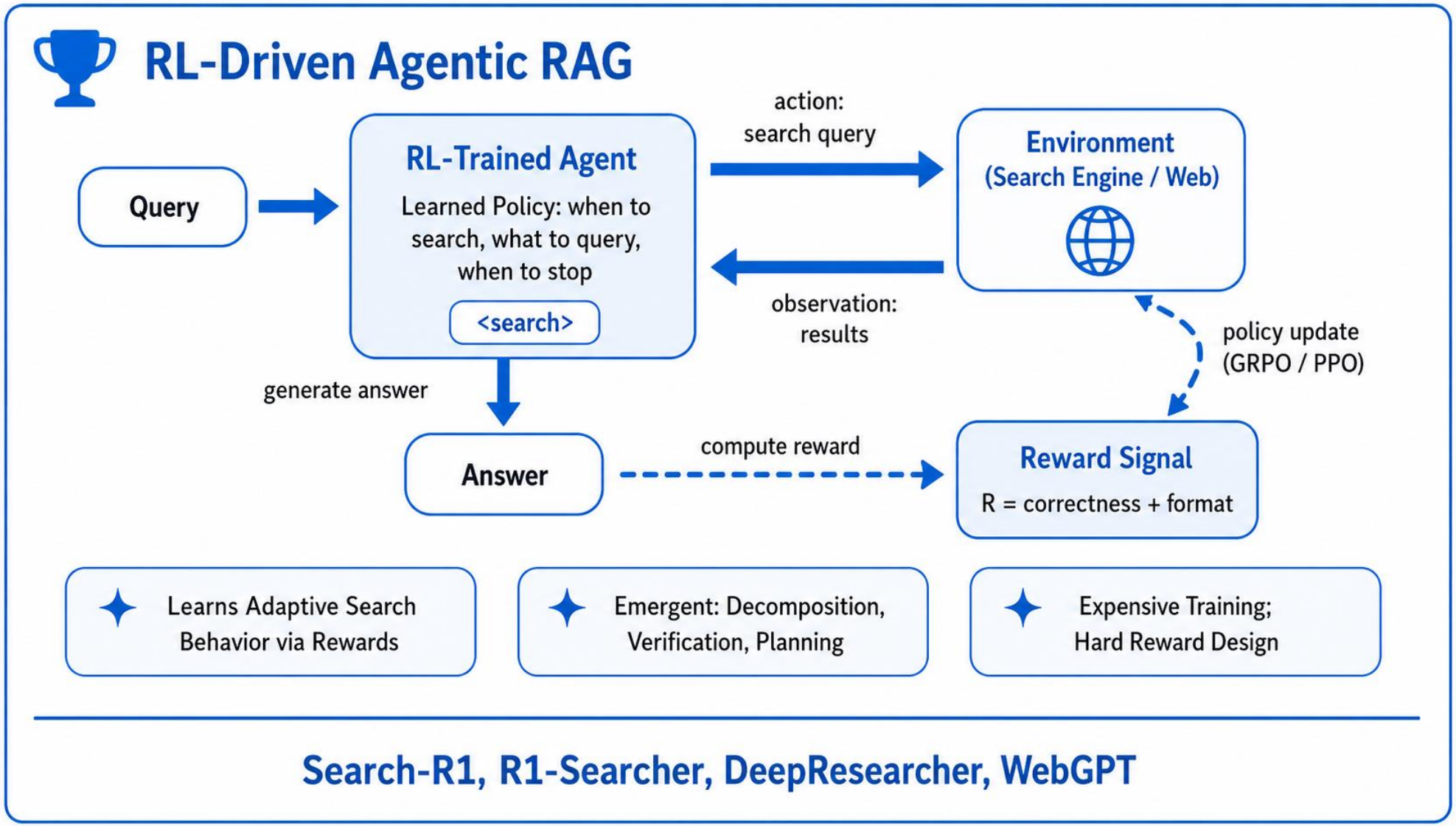
Wu et al. Agentic Reasoning: Reasoning LLMs with Tools for the Deep Research, arXiv, 2025

Multi-Agent Agentic RAG



Zhang et al. Chain of Agents: Large Language Models Collaborating on Long-Context Tasks, NeurIPS, 2024.

RL-Driven Agentic RAG



RL-Driven Agentic RAG

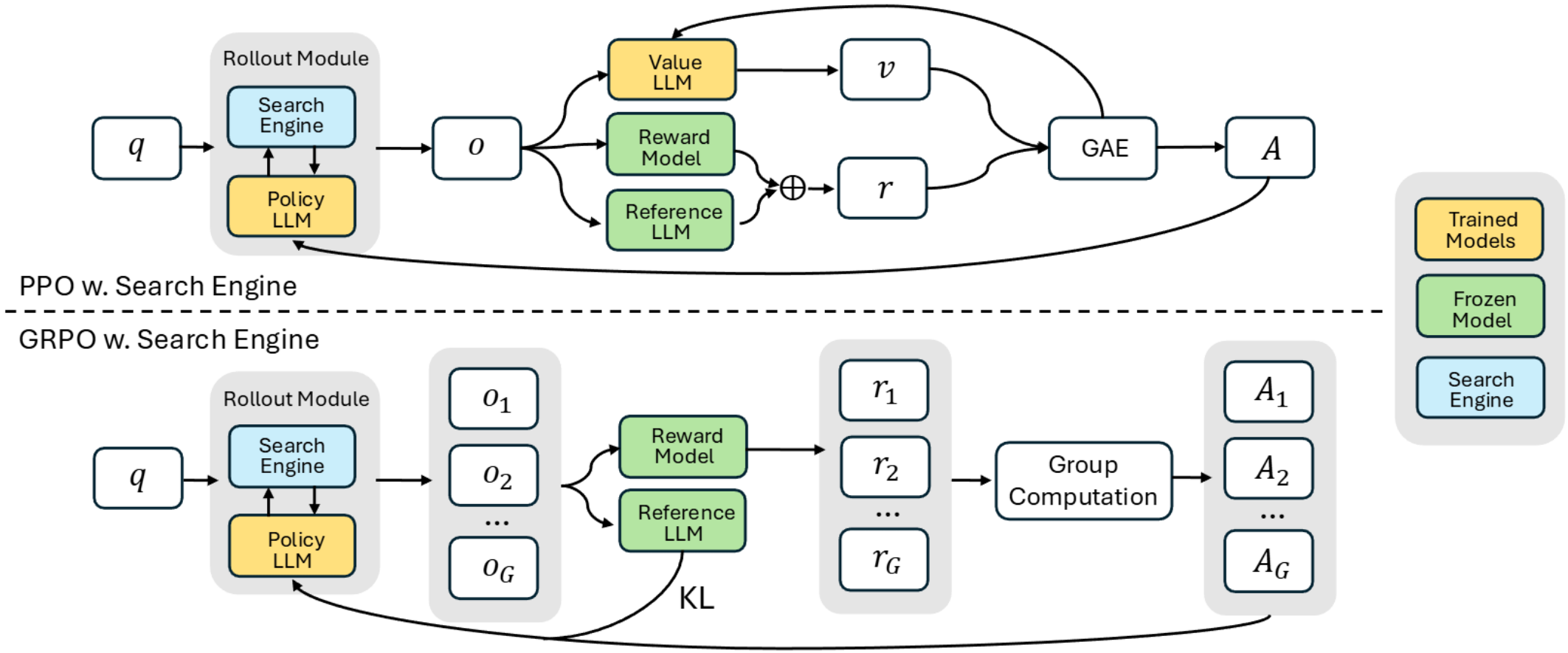
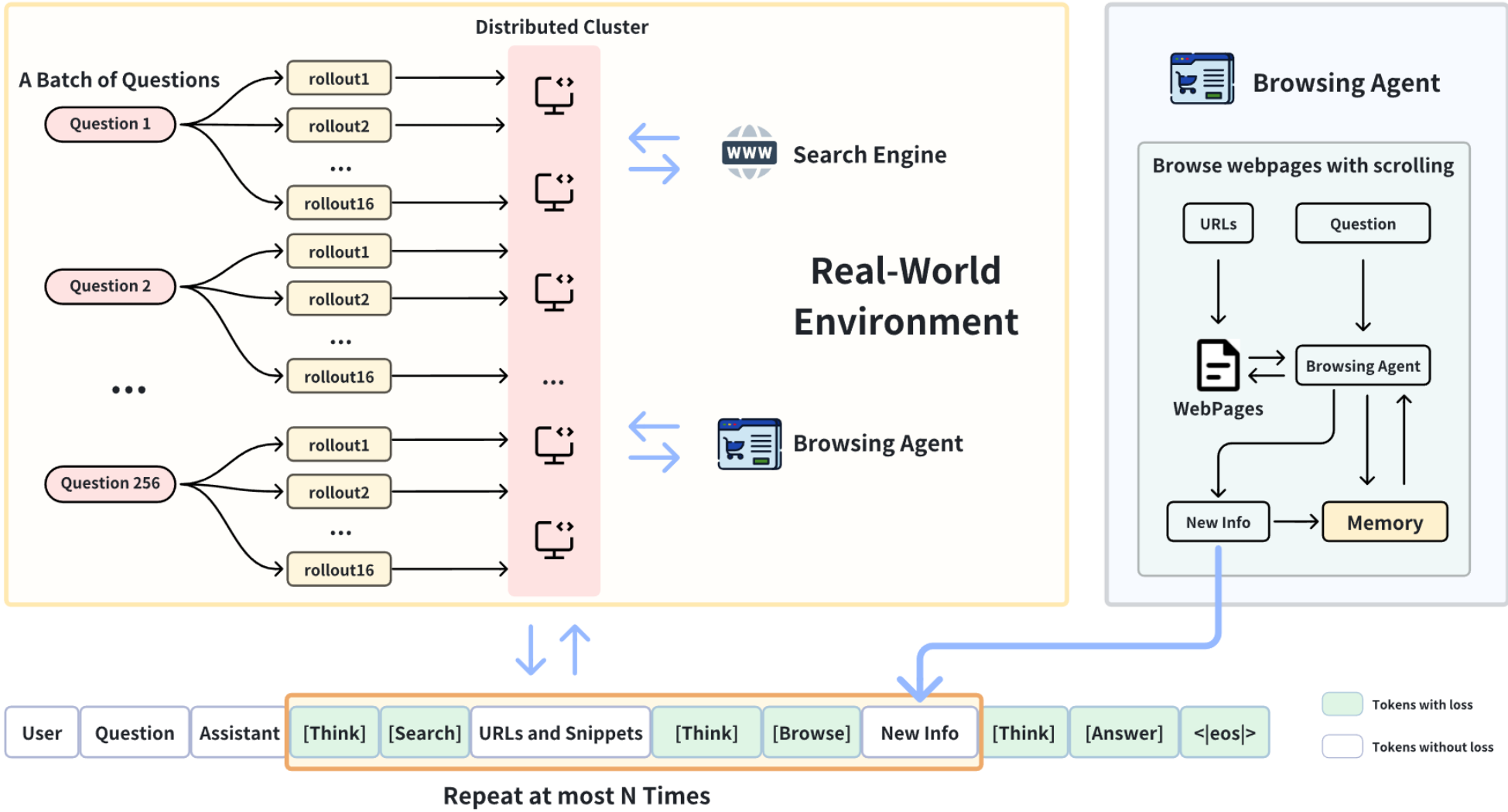


Figure 1: Demonstration of PPO and GRPO training with the search engine (SEARCH-R1). During the rollout, LLMs can conduct multi-turn interactions with the search engine.

Jin et al. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning, arXiv, 2025.

RL-Driven Agentic RAG



Zheng et al. DeepResearcher: Scaling Deep Research via Reinforcement Learning in Real-World Environments, arXiv, 2025.

Tutorial Outline

- **Part 1: Introduction** of Retrieval Augmented Large Foundation Models (RA-LFMs) (Dr. Wenqi Fan)
- **Part 2: Architecture** of RA-LFMs and Main Modules (Xu Yuan)
- **Part 3: Learning Approach** of RA-LFMs (Chengliang Liu)
- **Part 4: Agentic RAG** (Chengliang Liu)
- **Part 5: Applications of RA-LFMs (Chun-Hin Chan)**
- **Part 6: Challenges and Future Directions** of RA-LFMs (Dr. Wenqi Fan)
- **Part 7: Q&A**

Website of this tutorial
Check out the slides and more information!

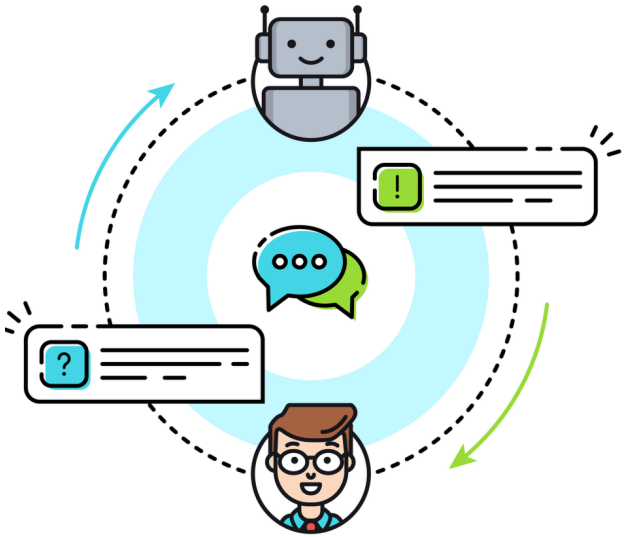
PART 4: Application of RA-LFMs

Presenter
Chun-Hin Chan
HK PolyU

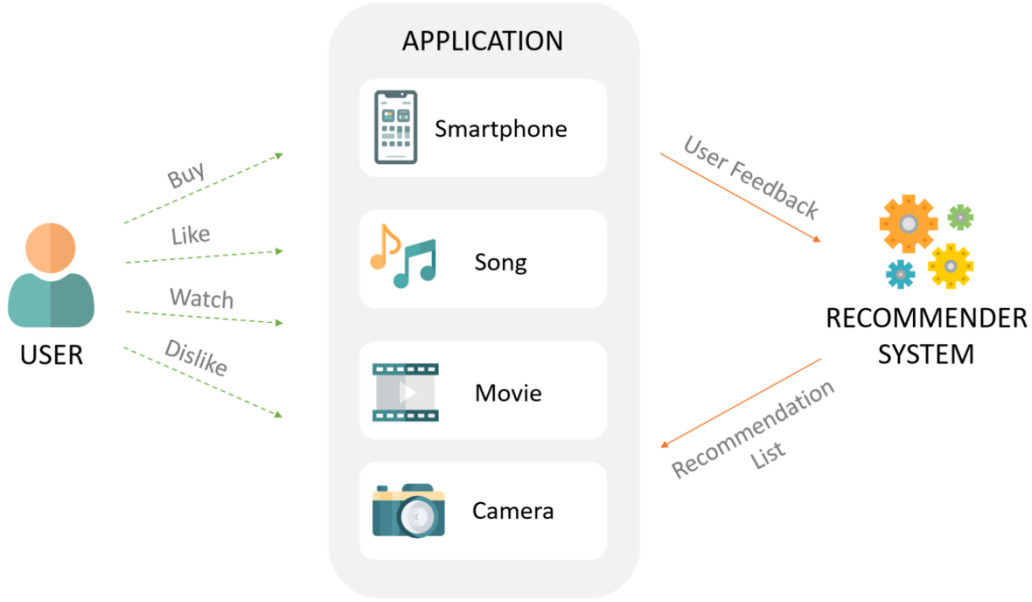
- **Preliminary applications in NLP**
- **Downstream tasks**
- **Domain-specific applications**

RA-LFM Applications

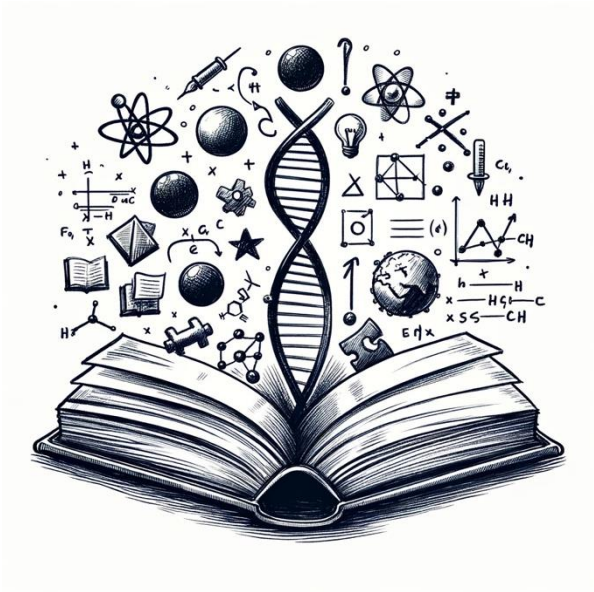
- **Various applications**



Chatbots



Recommendation



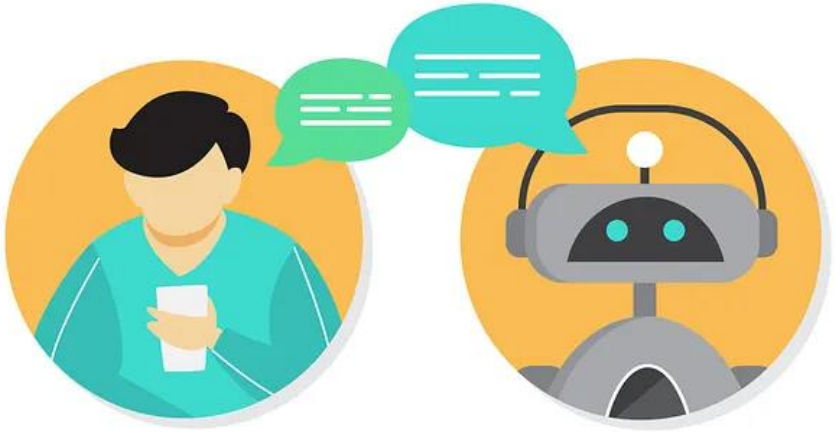
AI for Science

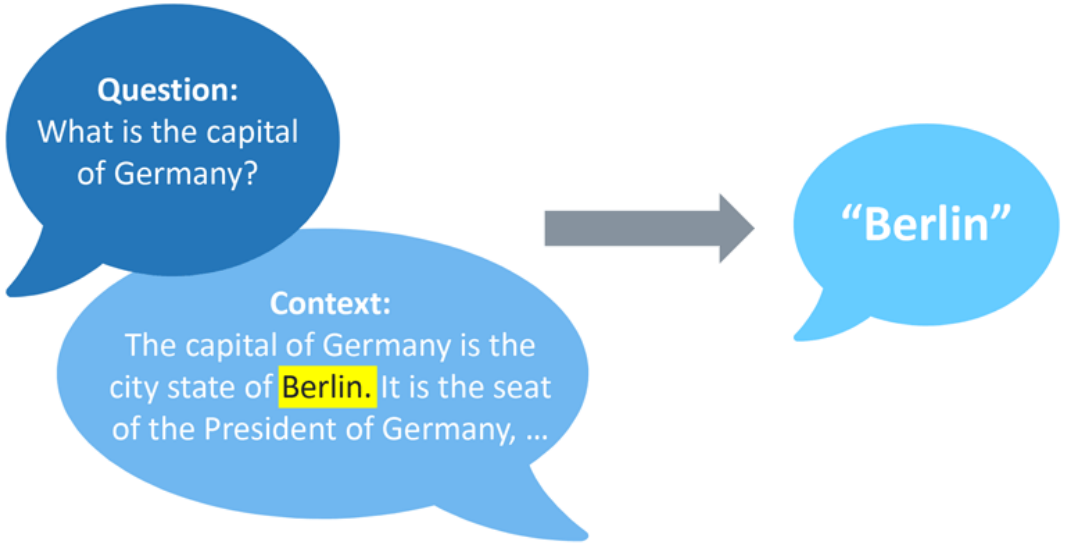
...

Ferreira, Diana, et al., 2020. “Recommendation System Using Autoencoders”. Applied Sciences 2020.
https://www.intelli-science.com/p/large-science-models-in-2024-hype

RA-LFM Applications: QA Systems

- **QA Systems**
  - QA systems aim to provide **precise answers** to user's queries





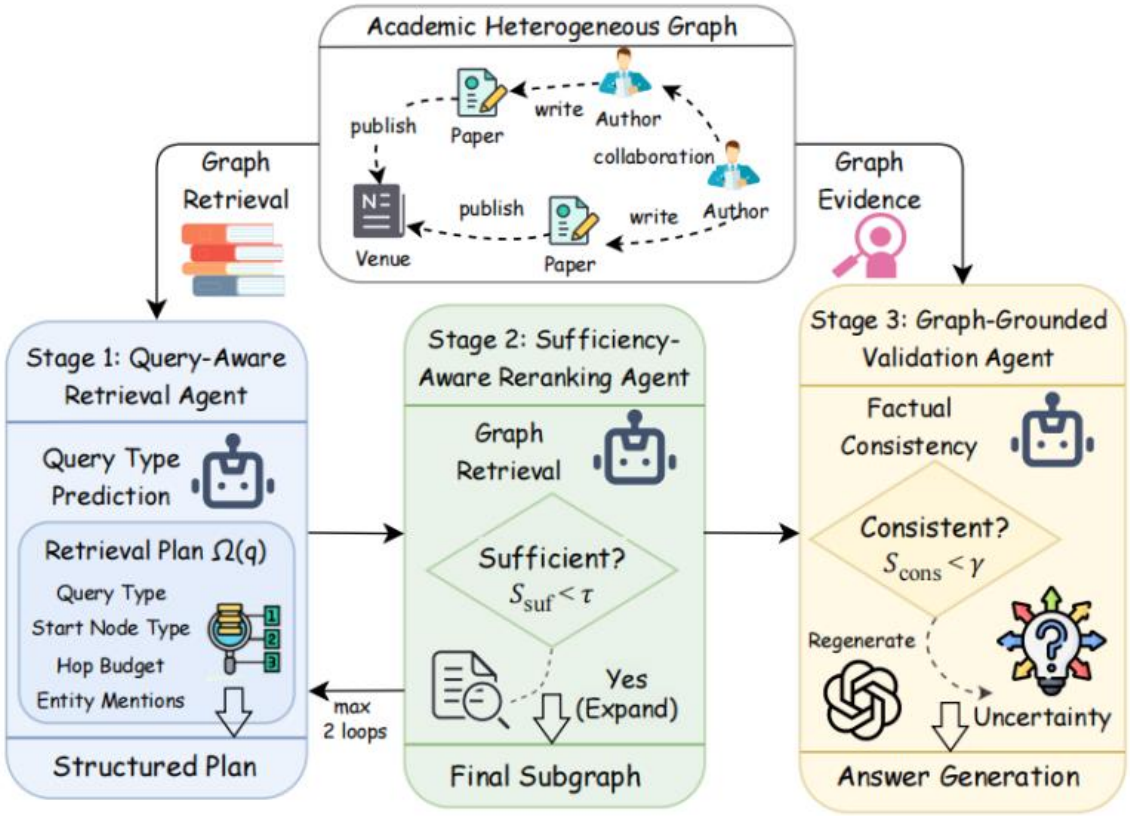
https://medium.com/mindmeld/question-answering-for-conversational-interfaces-5d6af5aff1e5
https://blog.seeburger.com/natural-language-question-answering-systems-get-quick-answers-to-concrete-questions/

RA-LFM Applications: QA Systems

- **QA Systems**
  - Challenges:
    - Fail to adapt to **varying query complexity**
    - **Incomplete or misaligned evidence** due to lack of sufficiency evaluation
    - **No structured verification** against graph facts

RA-LFM Applications: QA Systems

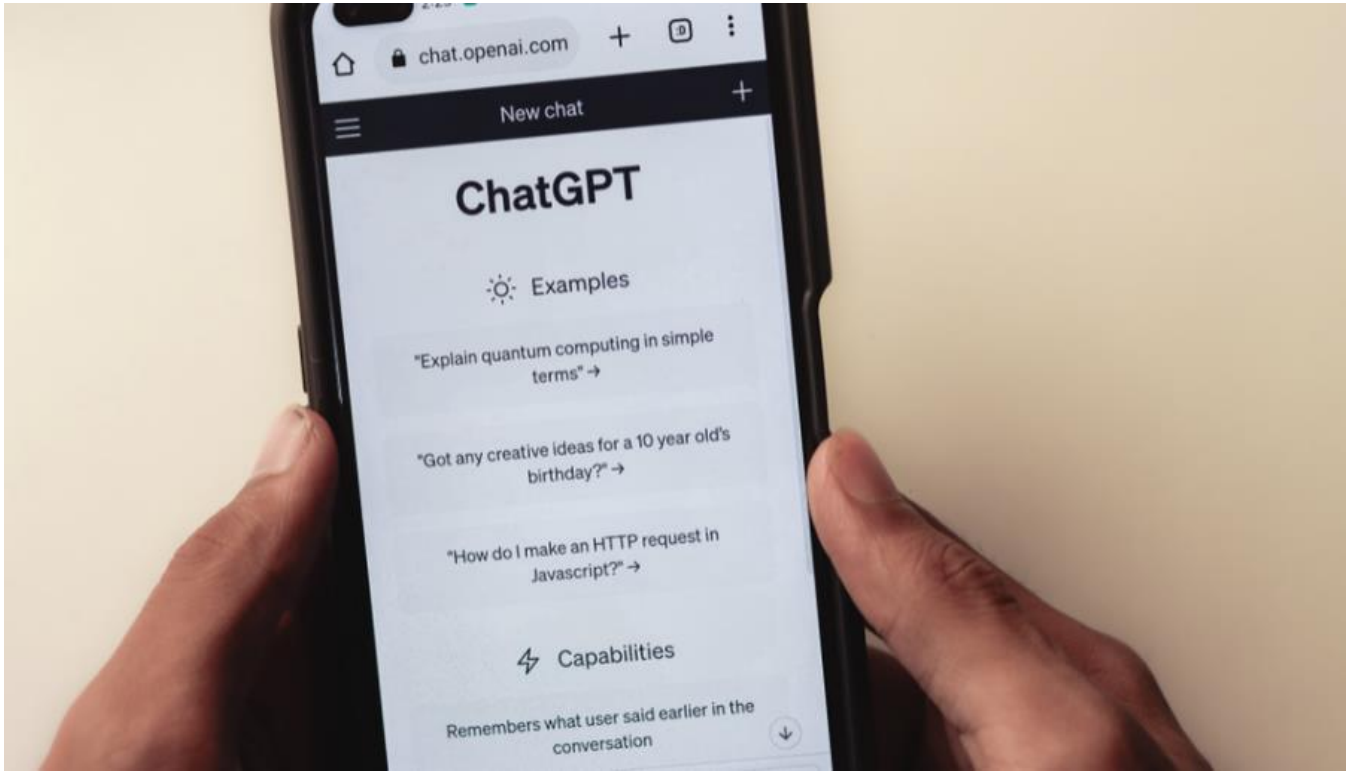
- **Retrieves from academic heterogeneous graph for Academic QA**
  - Agent-Enhanced Heterogeneous Graph RAG for Academic Question Answering



Jia et al., 2026. "Agent-Enhanced Heterogeneous Graph RAG for Academic Question Answering" WWW 2026.

RA-LFM Applications: Chatbots

- **Chatbots**
  - Chatbot interacts with users in a **natural** & **conversational** manner



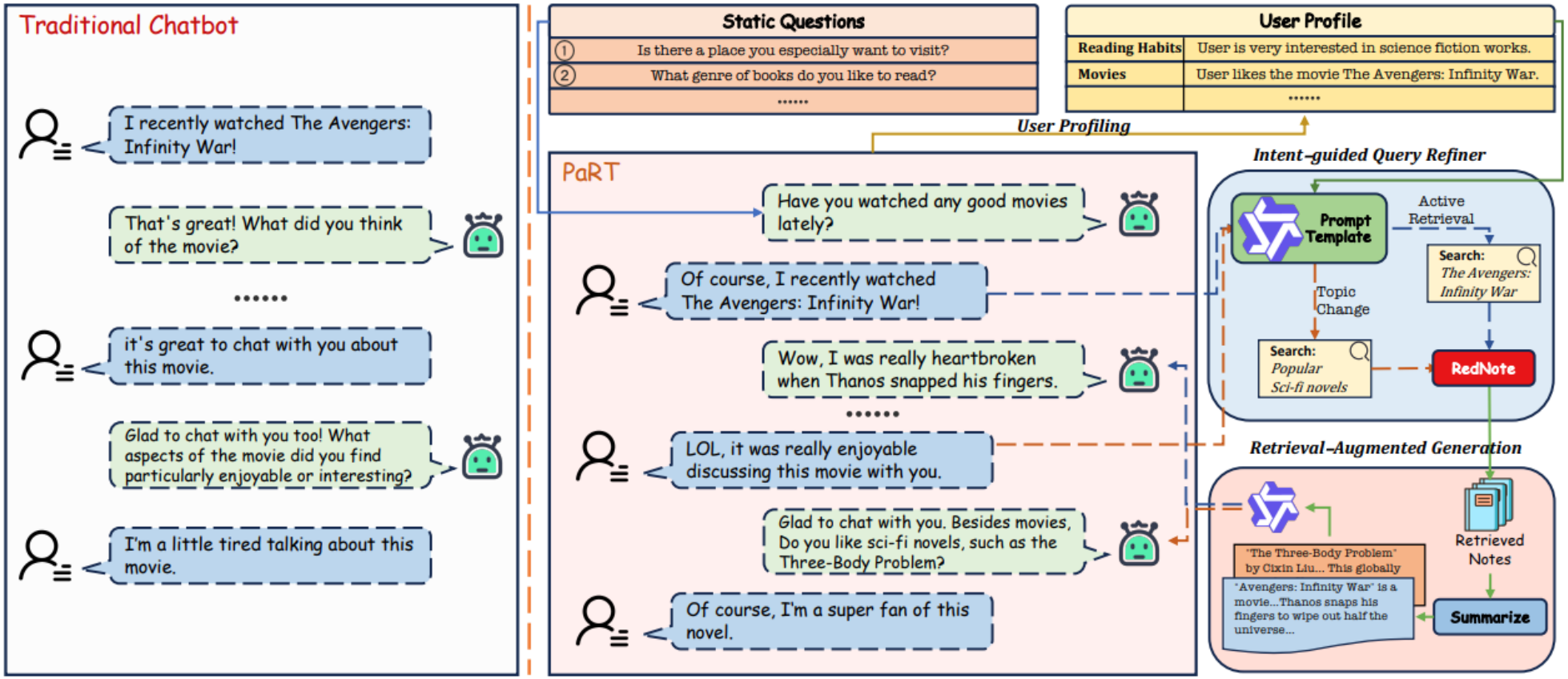
https://www.forbes.com/advisor/in/business/software/what-is-a-chatbot/

RA-LFM Applications: Chatbots

- **Chatbots**
  - Challenges:
    - **Neglect active engagement** with users
    - Constrain to **limited depth** & **natural extension of conversation**
    - **Short interaction** duration

RA-LFM Applications: Chatbots

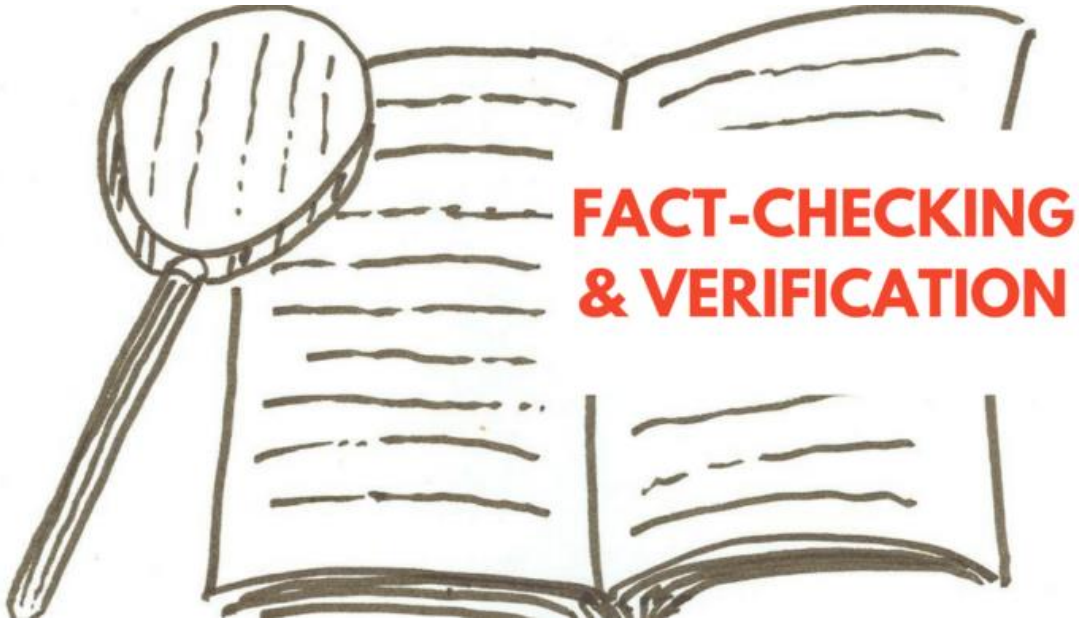
- **Retrieval from textual passages for Chatbots**
  - **P**ro**a**ctive social chatbots with personalized real-time **R**e**T**rieval (PaRT)



Niu et al., 2025. "PaRT: Enhancing Proactive Social Chatbots with Personalized Real-Time Retrieval". SIGIR 2025.

RA-LFM Applications: Fact Verification

- **Fact verification**
  - Fact Verification is a critical task in verifying the **accuracy** & **reliability** of information



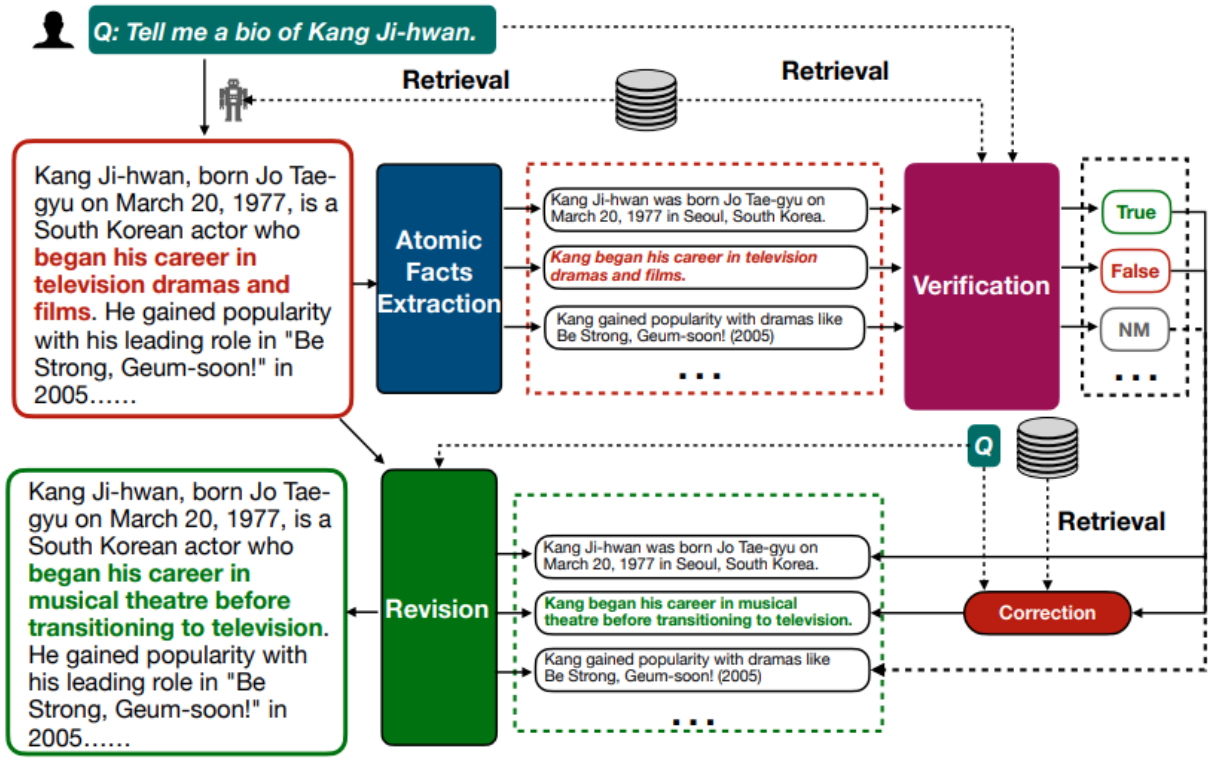
https://gijn.org/resource/fact-checking-verification/

RA-LFM Applications: Fact Verification

- **Fact verification**
  - Challenges:
    - LFMs have **limited access** and **manipulation** on knowledge
    - RAG does not guarantee **factual content**
    - Generate **factually incorrect output**

RA-LFM Applications: Fact Verification

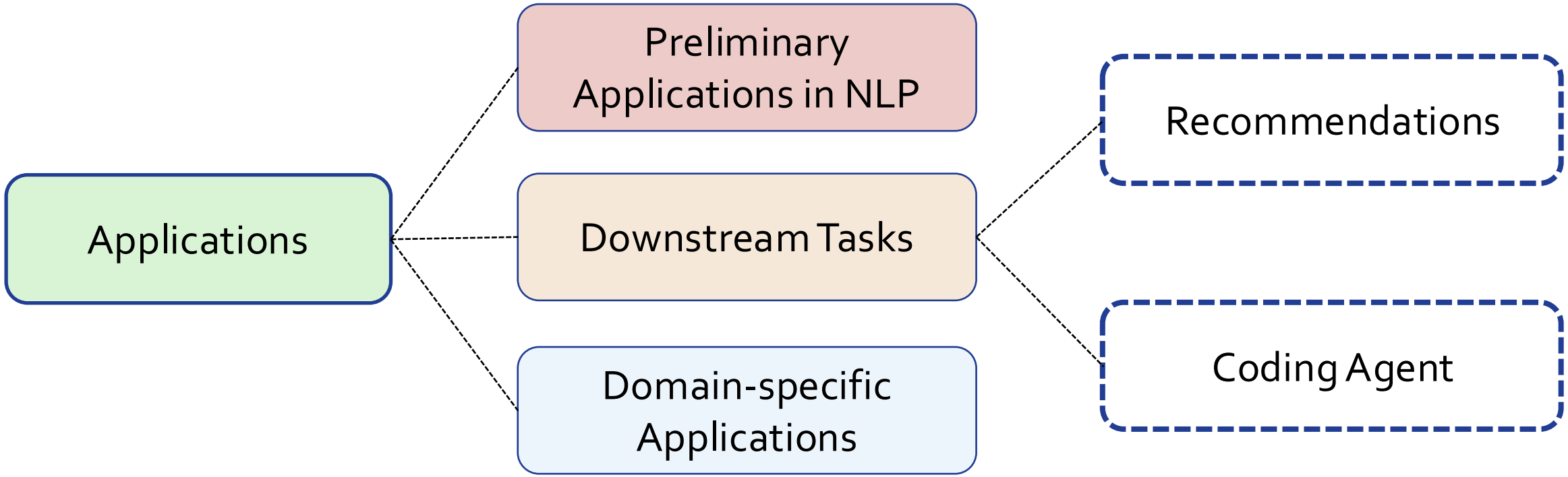
- **Retrieval from factual documents for Fact Verification**
  - **R**etrieval **A**ugmented **C**orrection (RAC)



Li & Flanigan, 2025. "RAC: Efficient LLM Factuality Correction with Retrieval Augmentation". EMNLP 2025.

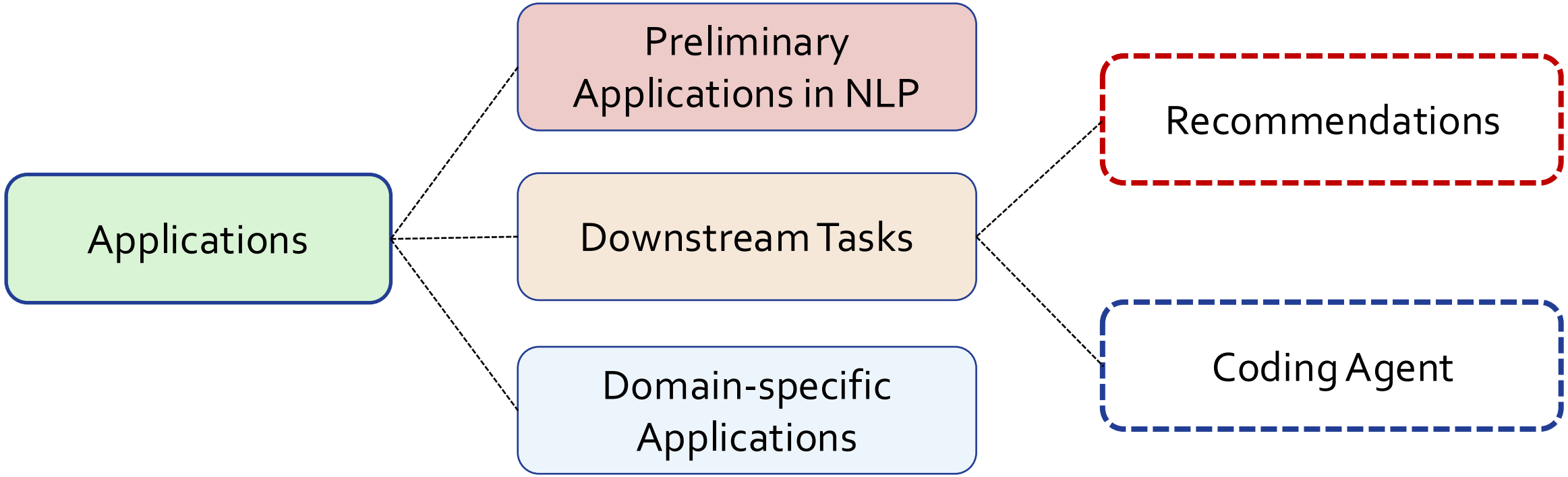
RA-LFM Applications: Downstream Tasks

- **Downstream tasks**



RA-LFM Applications: Recommendations

- **Recommendations**



RA-LFM Applications: Recommendations

- **Recommendations**
  - Recommendation has been widely applied in **online services**

YouTube

TikTok

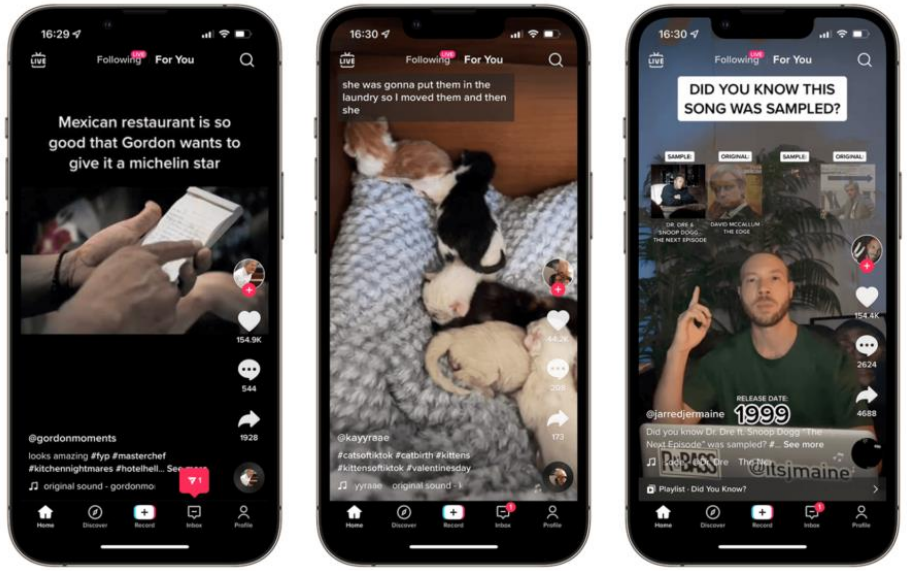
Google News

News/Video/Image Recommendation

TikTok's recommendation algorithm

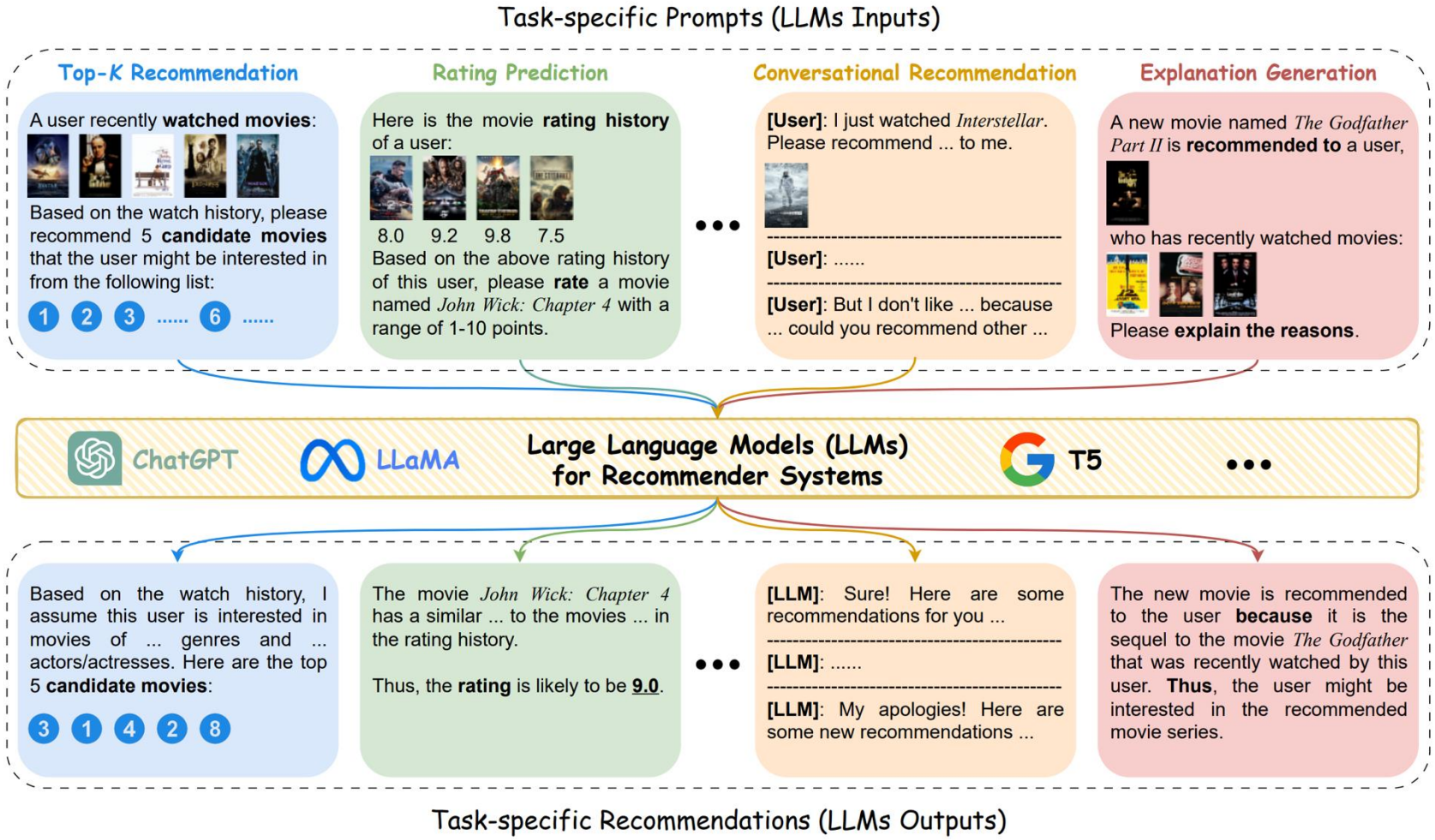
Top 10 Global Breakthrough Technologies in 2021

MIT Technology Review



RA-LFM Applications: Recommendations

- **LFMs in recommendations**



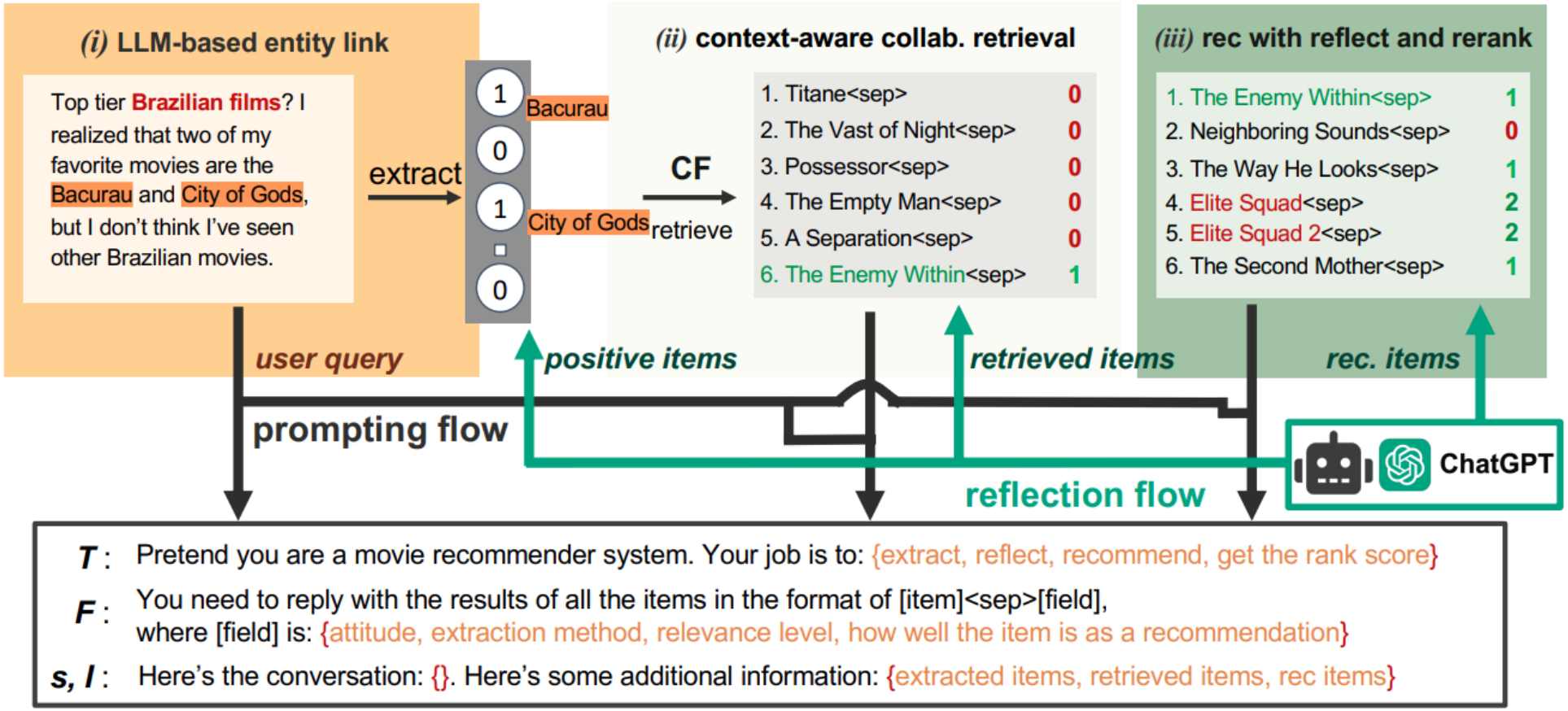
Zhao et al., 2024. "Recommender Systems in the Era of Large Language Models (LLMs)". TKDE 2024.

RA-LFM Applications: Recommendations

- **Recommendations**
  - Challenges:
    - Fall short in the effective usage of **collaborative filtering knowledge**
    - **Neglect structural relationships** in knowledge
    - Matching-based RAG unable to identify **useful information over noisy web data**

RA-LFM Applications: Recommendations

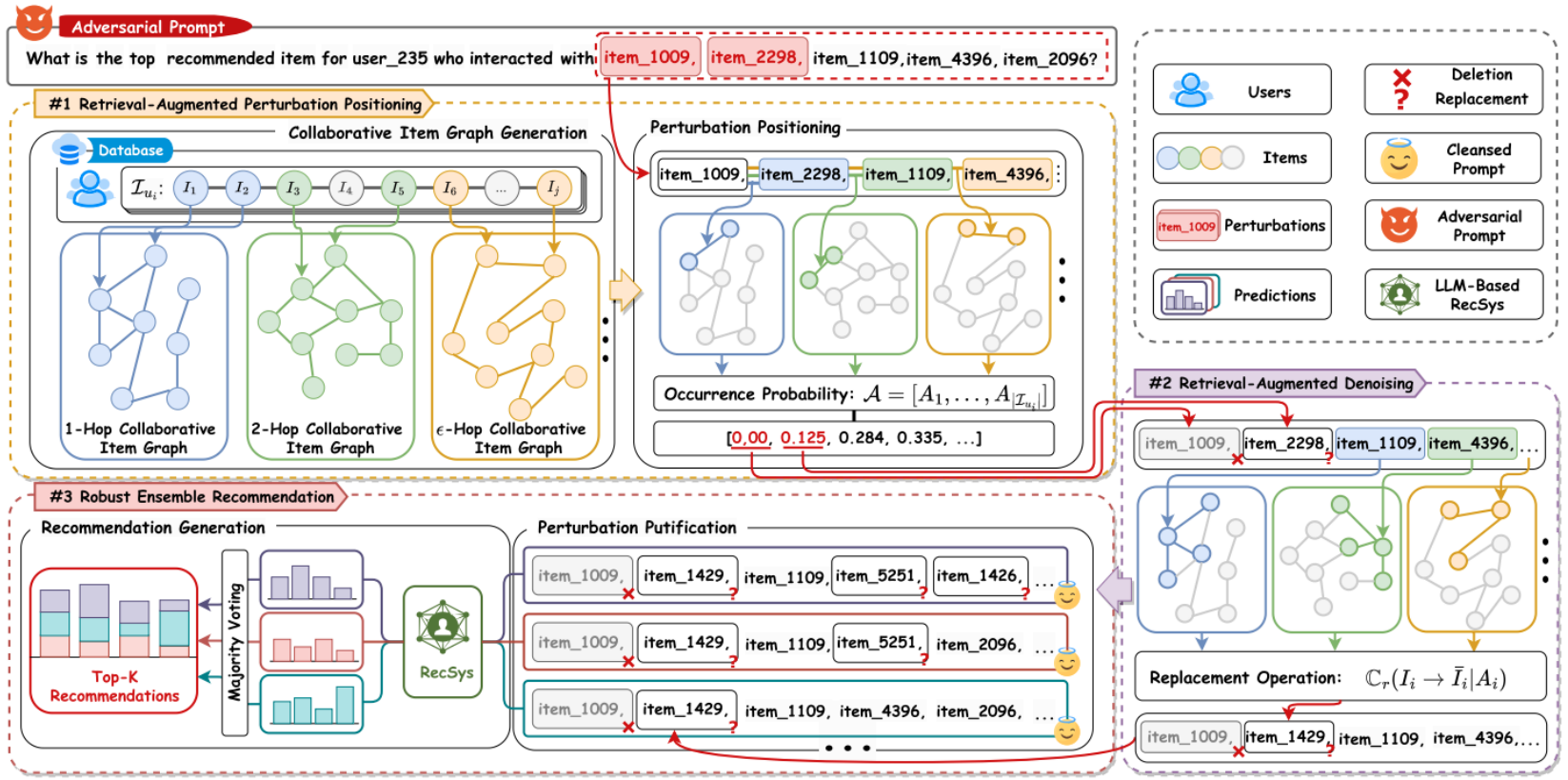
- **Retrieval from collaborative signals**
  - **C**ollaborative **R**etrieval **A**ugmented **G**eneration (CRAG)



Zhu et al., 2025. "Collaborative Retrieval for Large Language Model-based Conversational Recommender Systems". WWW 2025.

RA-LFM Applications: Recommendations

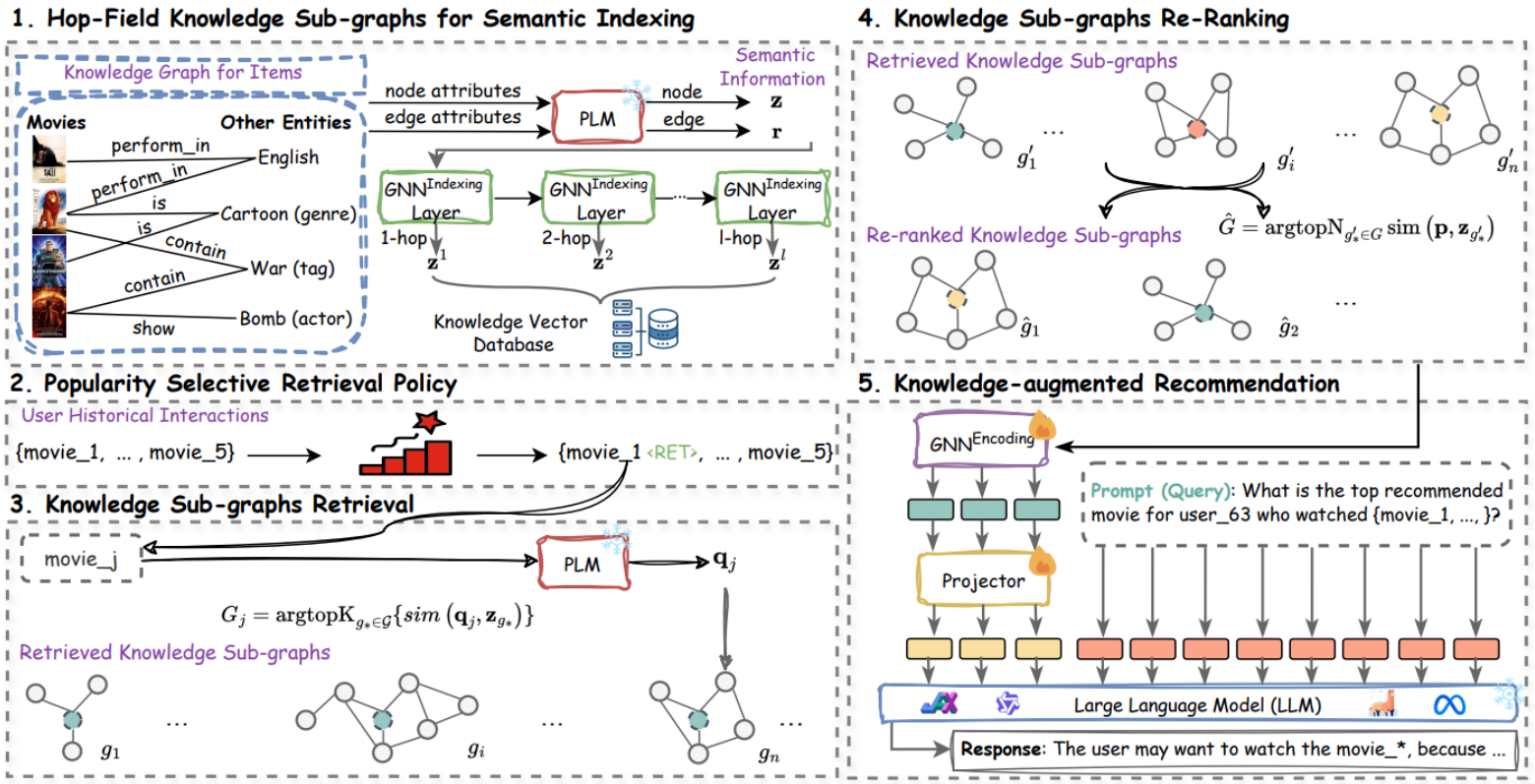
- **Retrieval from collaborative signals for Purification**
  - **RE**trieval-a**U**gumented pu**R**ifier for e**N**hancing robustness of LLM based RecSys (RETURN)



Ning et al., 2026. "Retrieval-Augmented Purifier for Robust LLM-Empowered Recommendation". TOIS 2026.

RA-LFM Applications: Recommendations

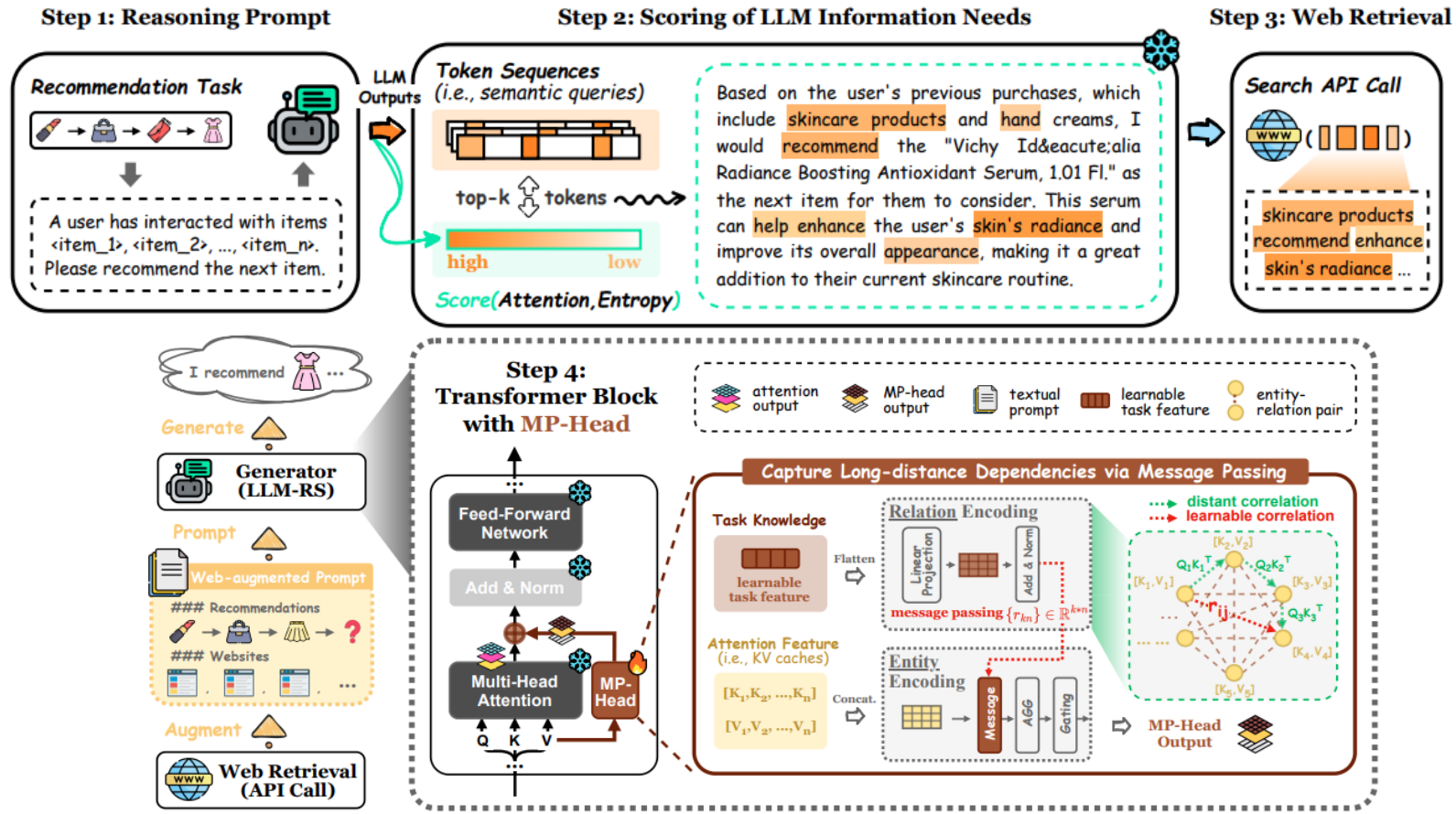
- **Retrieval from knowledge graph**
  - **K**nowledge graph **R**etrieval-**A**ugmented **G**eneration for LLM-based **Rec**ommendation (K-RagRec)



Wang et al., 2025. "Knowledge Graph Retrieval-Augmented Generation for LLM-based Recommendation". ACL 2025.

RA-LFM Applications: Recommendations

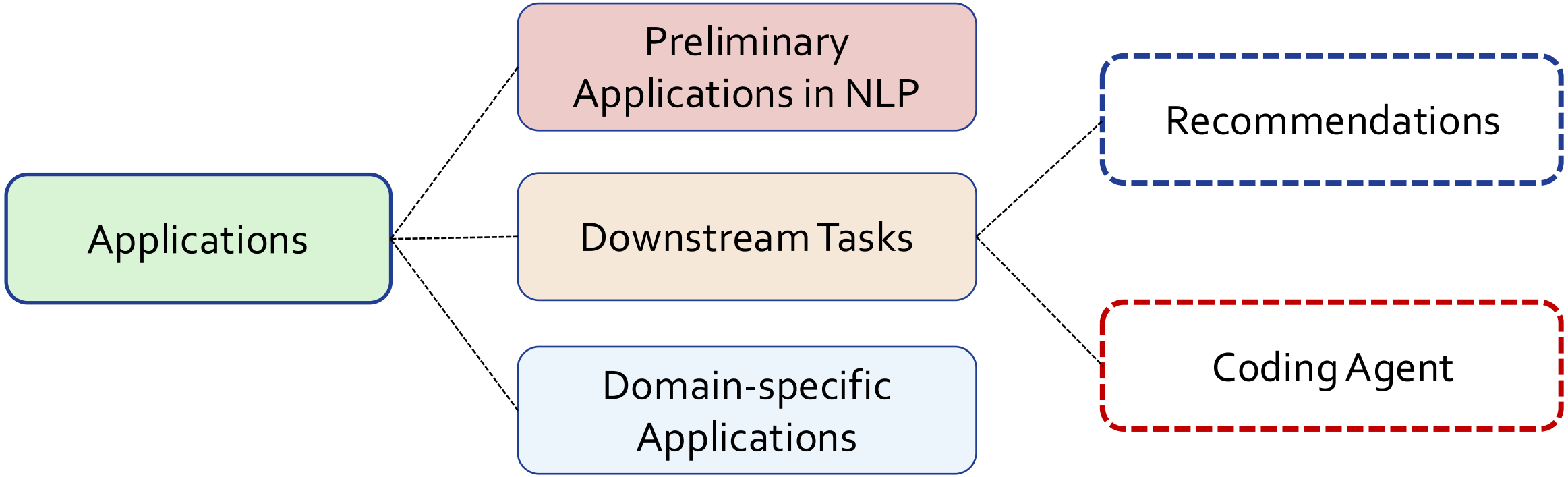
- **Retrieval from the web**
  - RAG from **Web** for enhancing LLM-based **Rec**ommendations (WebRec)



Zhao et al., 2025. "WebRec: Enhancing LLM-based Recommendations with Attention-guided RAG from Web". ArXiv 2025.

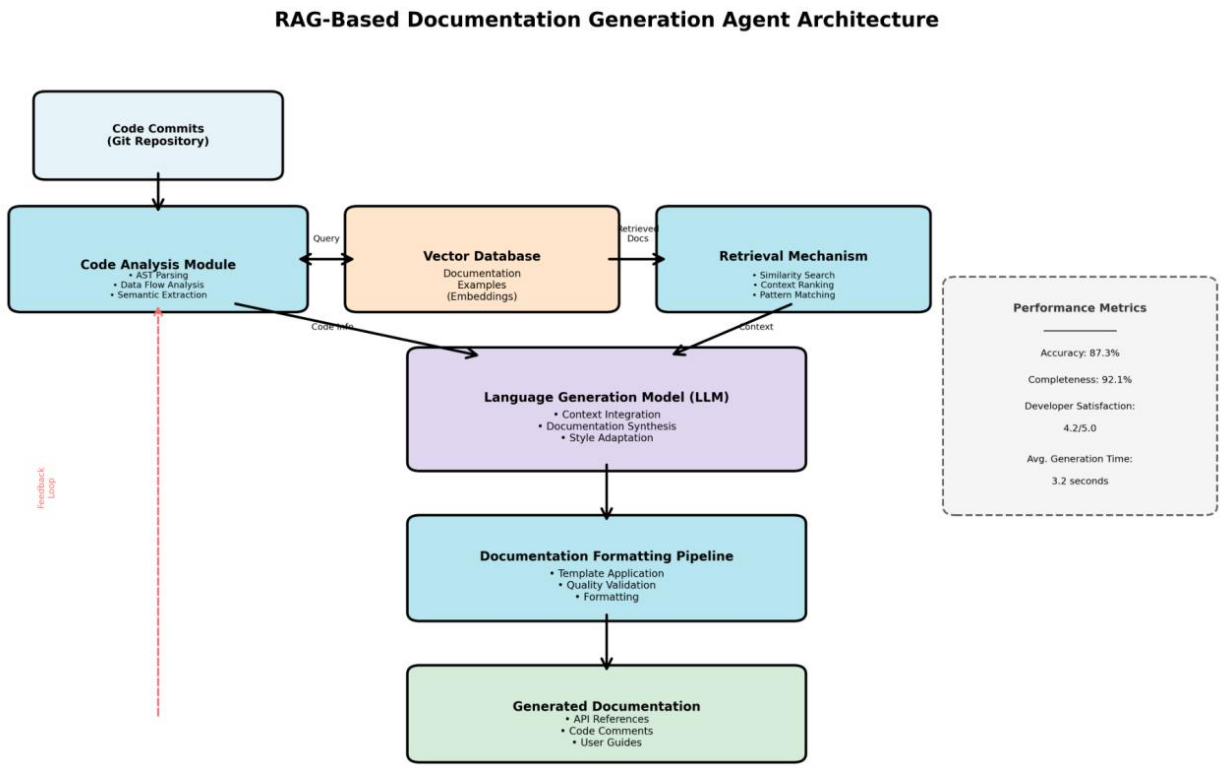
RA-LFM Applications: Coding Agent

- **Coding agent**



RA-LFM Applications: Coding Agent

- **Coding agent:**
  - Code generation
  - Code repair
  - Production deployment
  - …



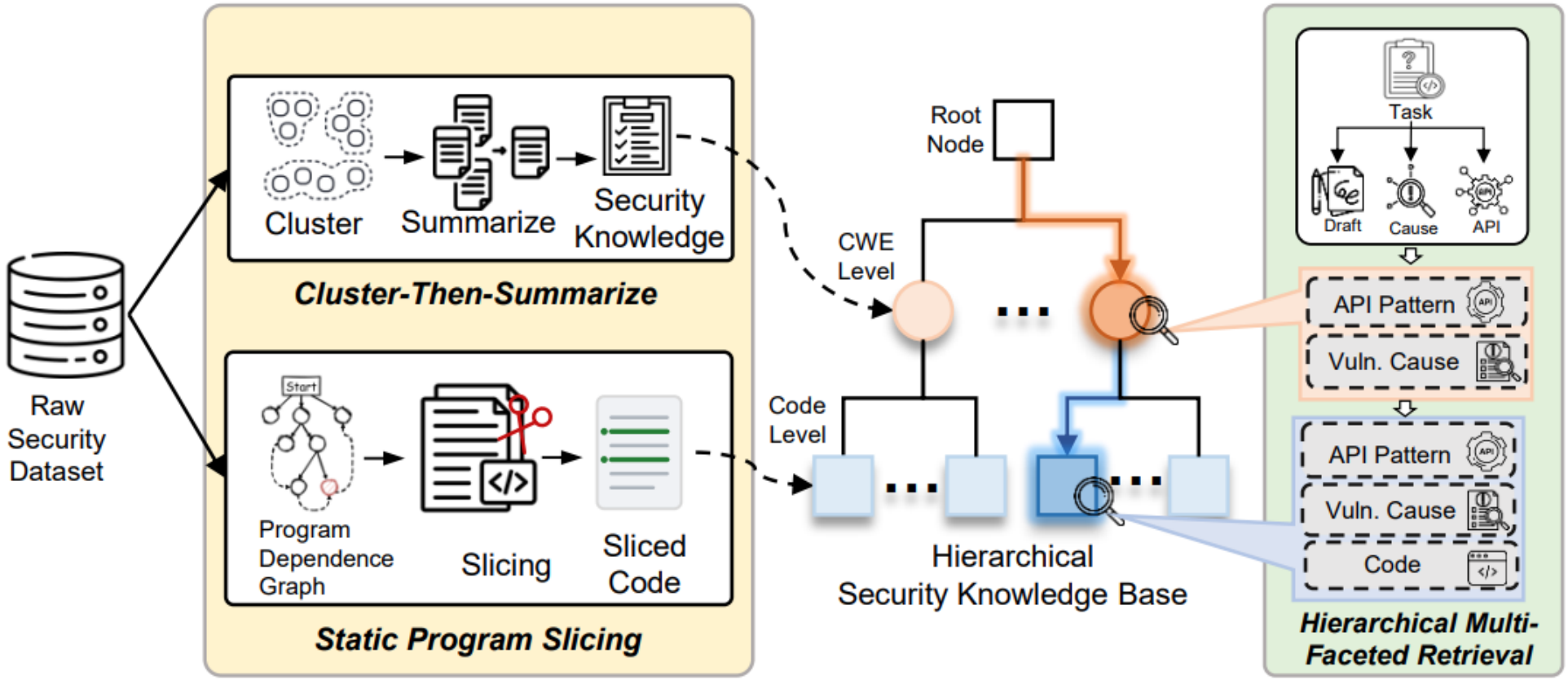
Zhao et al., 2025. "RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment". Frontiers in Artificial Intelligence Research 2025

RA-LFM Applications: Coding Agent

- **Coding Agent**
  - Challenges:
    - Struggles with the **noise of raw security-related documents**
    - Overlook the **key security semantics** implicitly embedded in task descriptions
    - Fail to capture the **structural intricacies of code**

RA-LFM Applications: Coding Agent

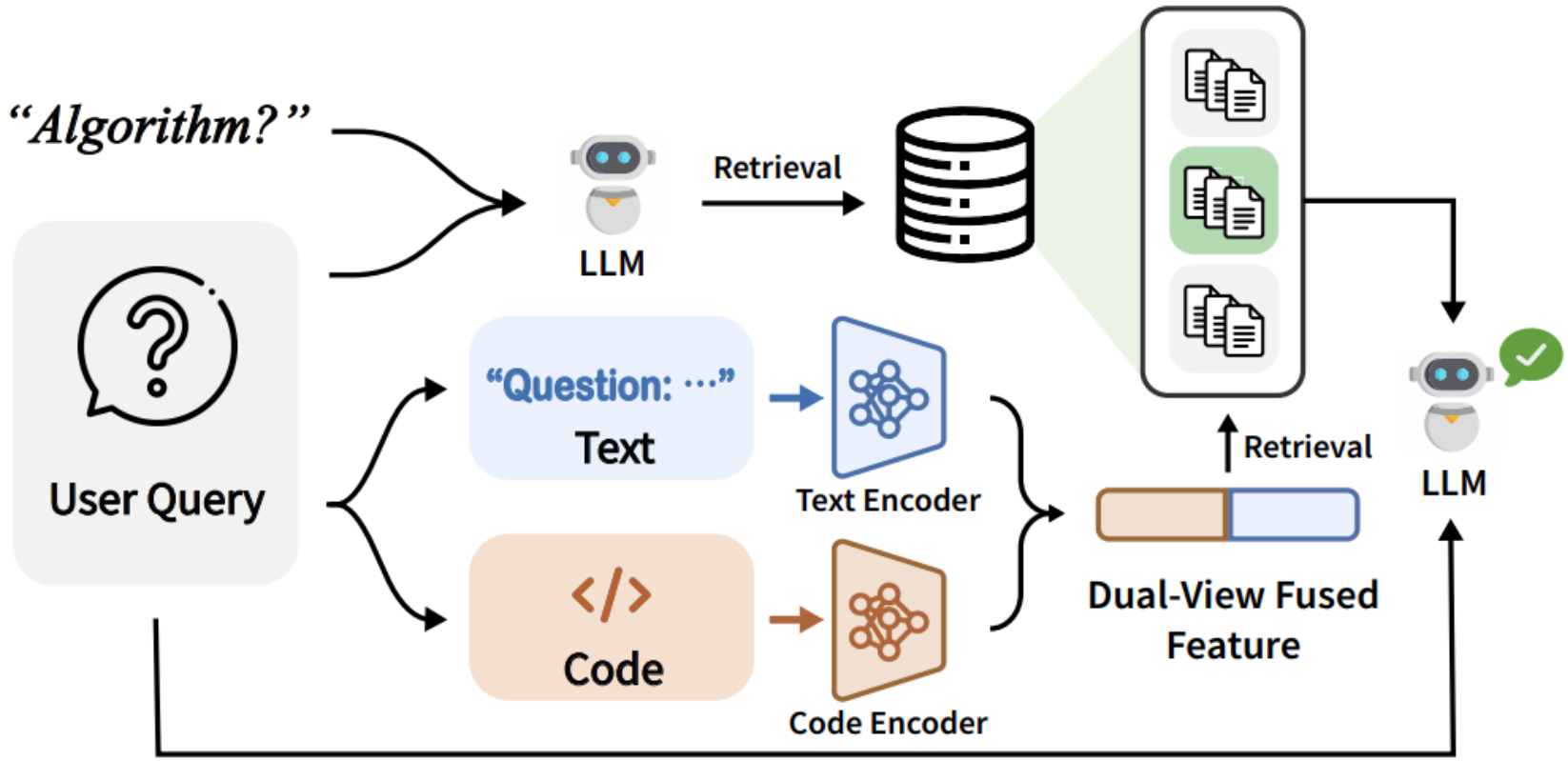
- **Retrieval from security knowledge base for Code Generation**
  - **RE**trieval-augmented **S**ecure **C**ode g**E**neration (RESCUE)



Shi & Zhang, 2026. "RESCUE: Retrieval Augmented Secure Code Generation". ICLR 2026.

RA-LFM Applications: Coding Agent

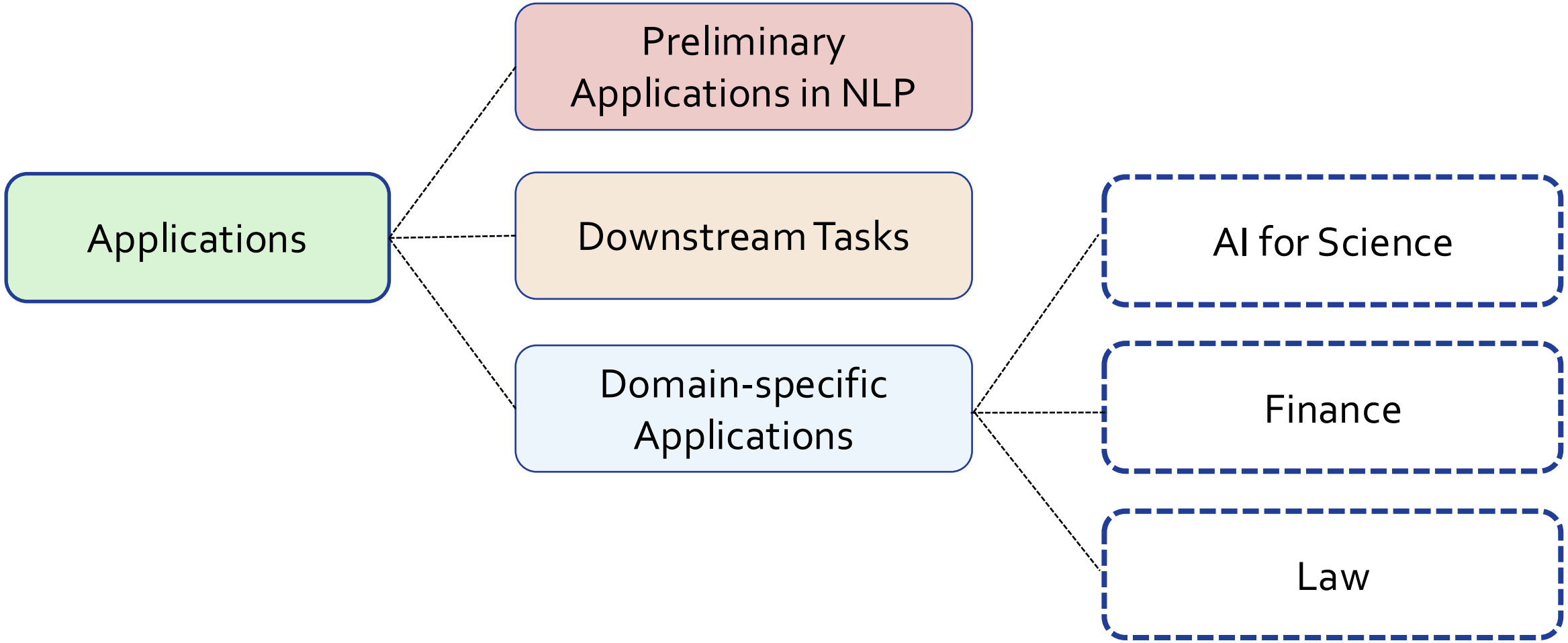
- **Retrieval from algorithm-specific knowledge corpus for Code Repair**
  - Improving LLM-based **Code Re**pair with Fine-Grained Retrieval-Augmented Generation (ReCode)



Zhao et al., 2025. "ReCode: Improving LLM-based Code Repair with Fine-Grained Retrieval-Augmented Generation". CIKM 2025.

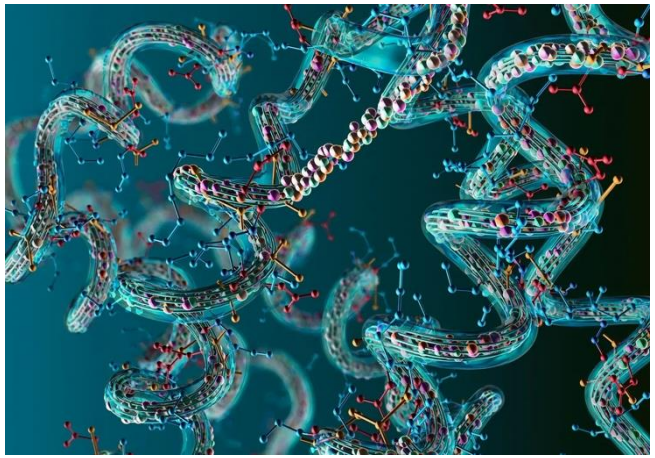
RA-LFM Applications: Domain-specific Applications

- **Domain-specific applications**



RA-LFM Applications: AI for Science

- **AI for science**
  - Molecules
  - Protein
  - …



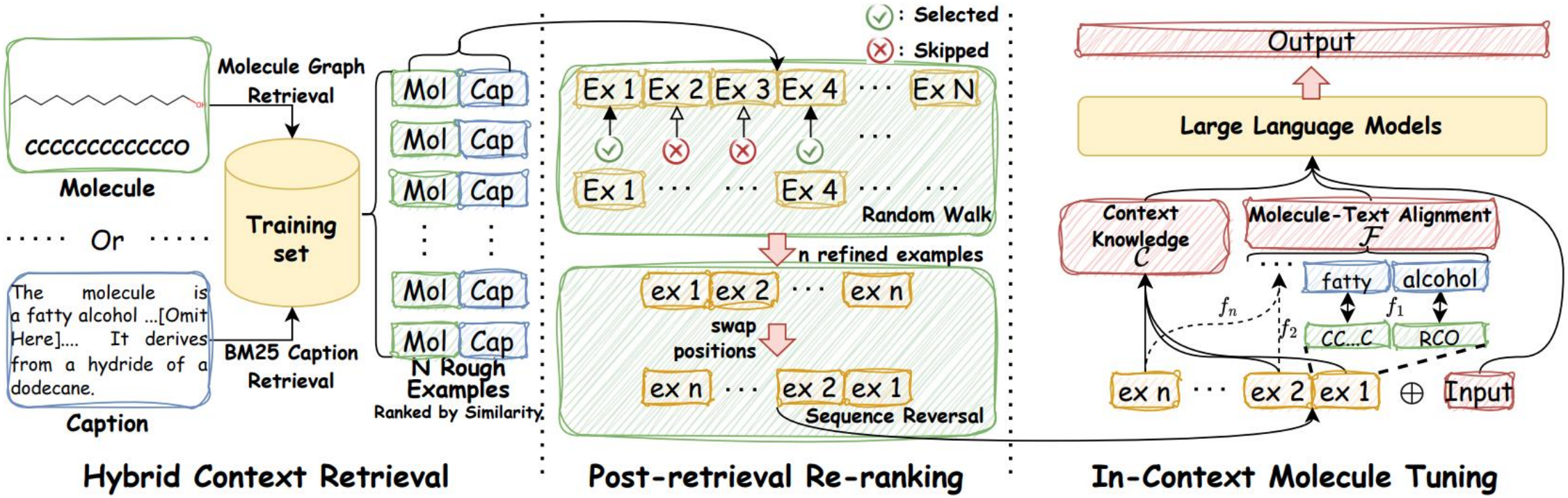
https://www.quantamagazine.org/how-ai-revolutionized-protein-science-but-didnt-end-it-20240626/

RA-LFM Applications: AI for Science

- **AI for science**
  - Challenges:
    - High-quality chemical datasets are **rare**
    - **Costly to finetune** on chemical-specific tasks
    - Fail to align **protein knowledge graphs** with **biological task specific data**

RA-LFM Applications: AI for Science

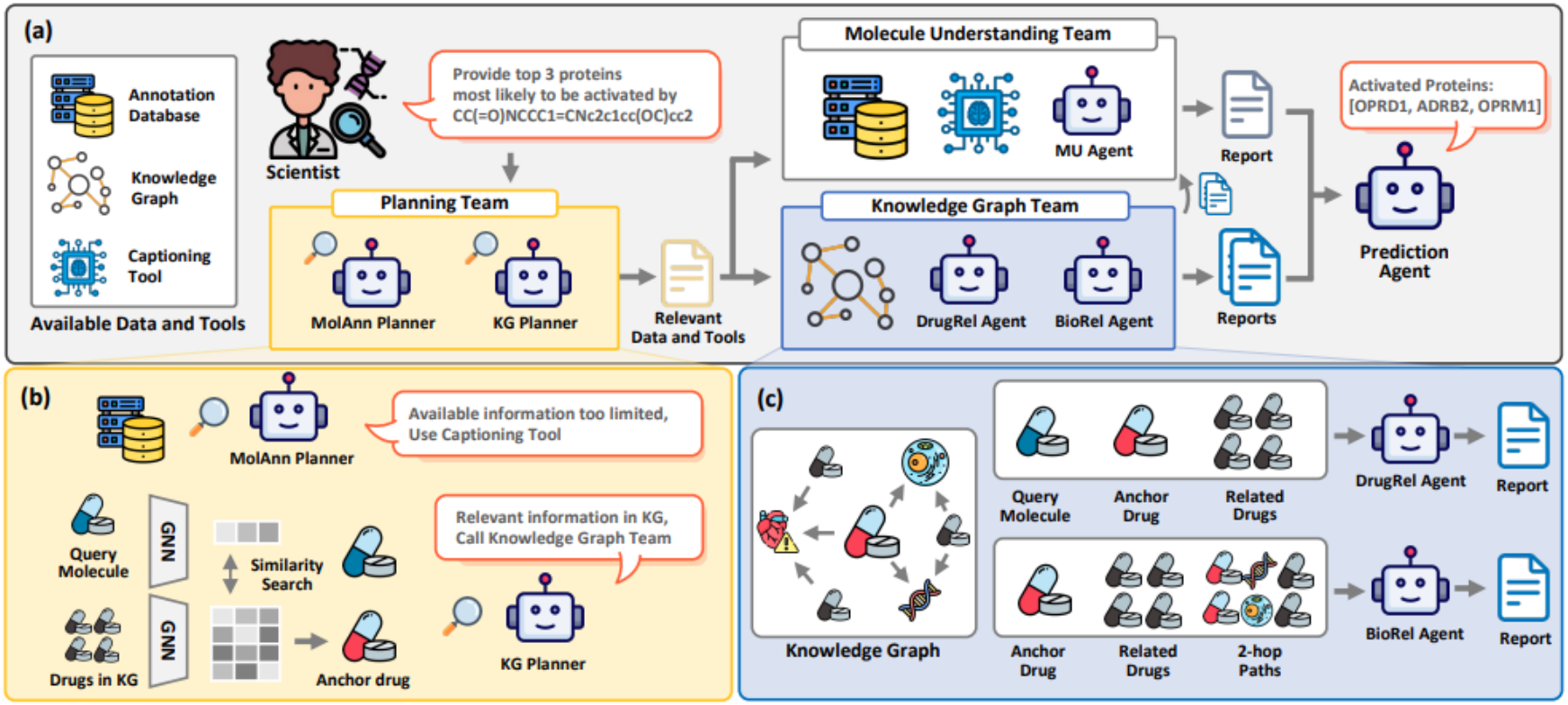
- **Retrieval from the molecule graph & caption for Molecules Learning**
  - **I**n-**C**ontext **M**olecule **A**daptation (ICMA)



Li et al., 2025. "Large Language Models are In-Context Molecule Learners". TKDE 2025.

RA-LFM Applications: AI for Science

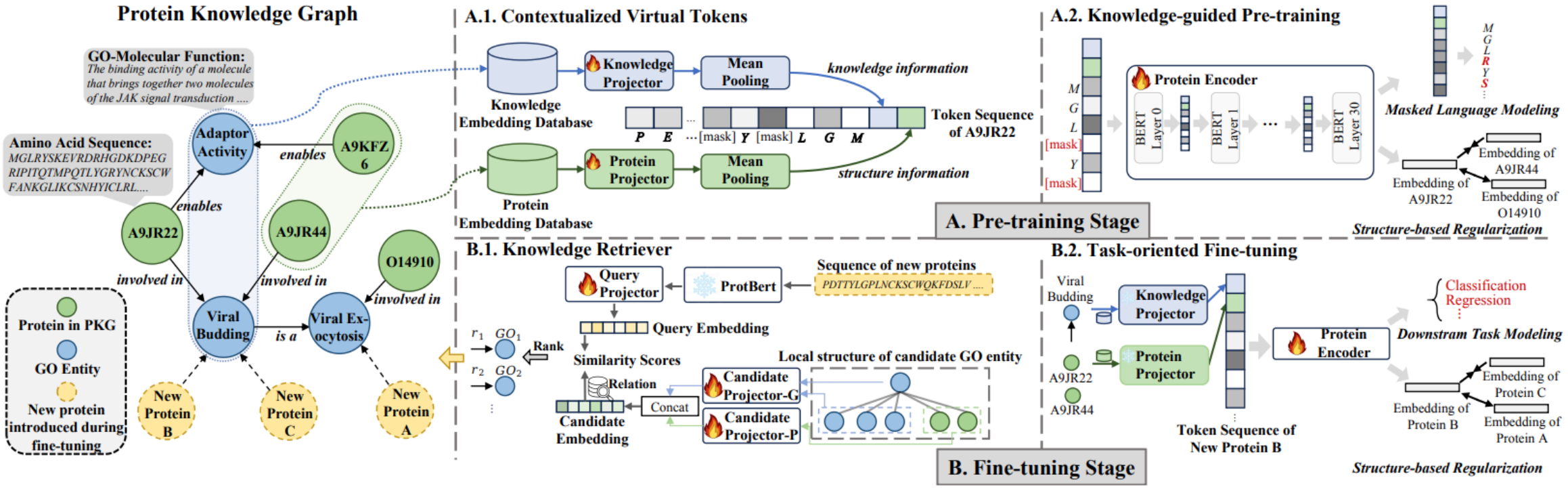
- **Retrieval from knowledge graph & molecule annotation databases for Drug Discovery**
  - **C**ollaborative framework of **LLM** **A**gents for **D**rug **D**iscovery (CLADD)



Lee et al., 2025. "RAG-Enhanced Collaborative LLM Agents for Drug Discovery". AAAI 2026.

RA-LFM Applications: AI for Science

- **Retrieval from protein knowledge graph for Protein Encoding**
  - **K**nowledge-**a**ware **r**etrieval **a**ugmented protein language model (Kara)



Zhang et al., 2025. "Retrieval-Augmented Language Model for Knowledge-aware Protein Encoding". ICML 2025.

RA-LFM Applications: Finance

- **Finance**
  - Financial question answering
  - Financial decision-making
  - …

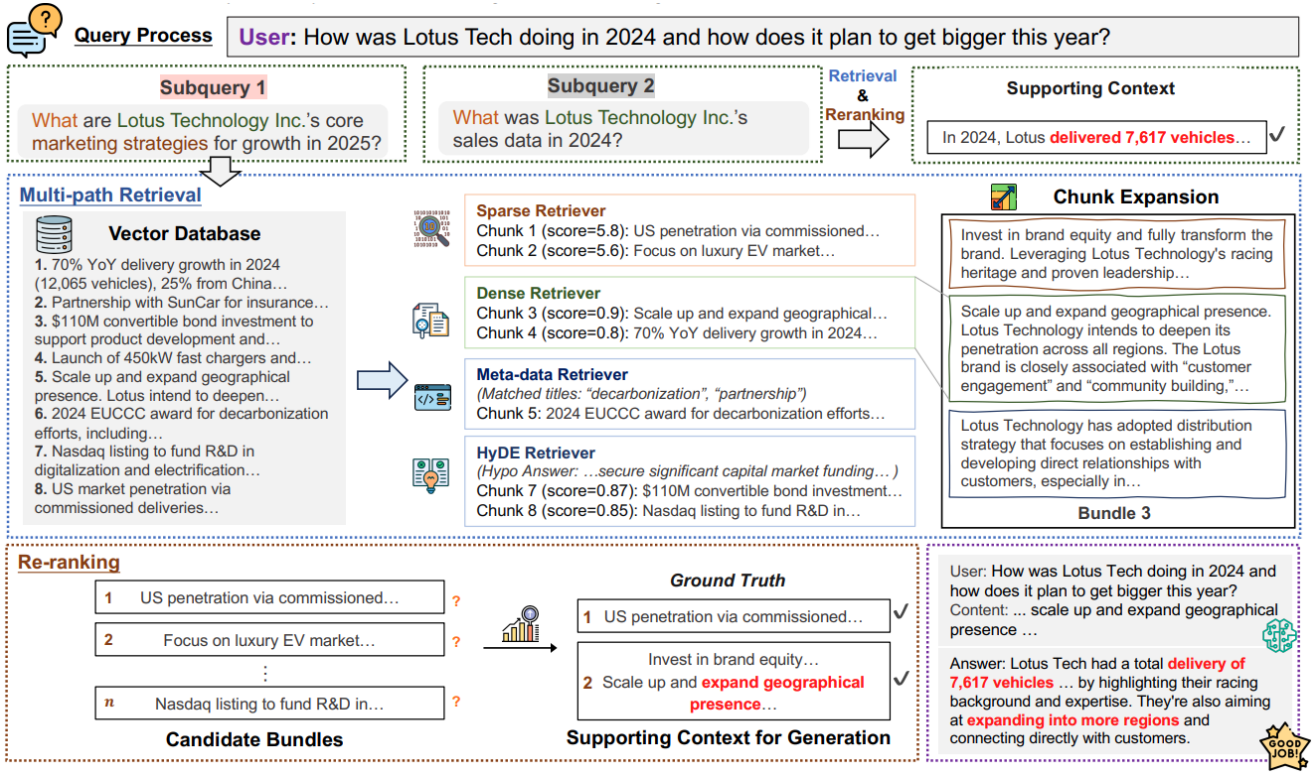
https://iconnect.isenberg.umass.edu/blog/2024/10/25/ai-in-finance-banking-11-ways-its-changing-the-industry/

RA-LFM Applications: Finance

- **Finance**
  - Challenges:
    - Financial regulations continuously **evolve**
    - Existing financial datasets offer **limited data modalities**
    - Fail to incorporate **temporal information**

RA-LFM Applications: Finance

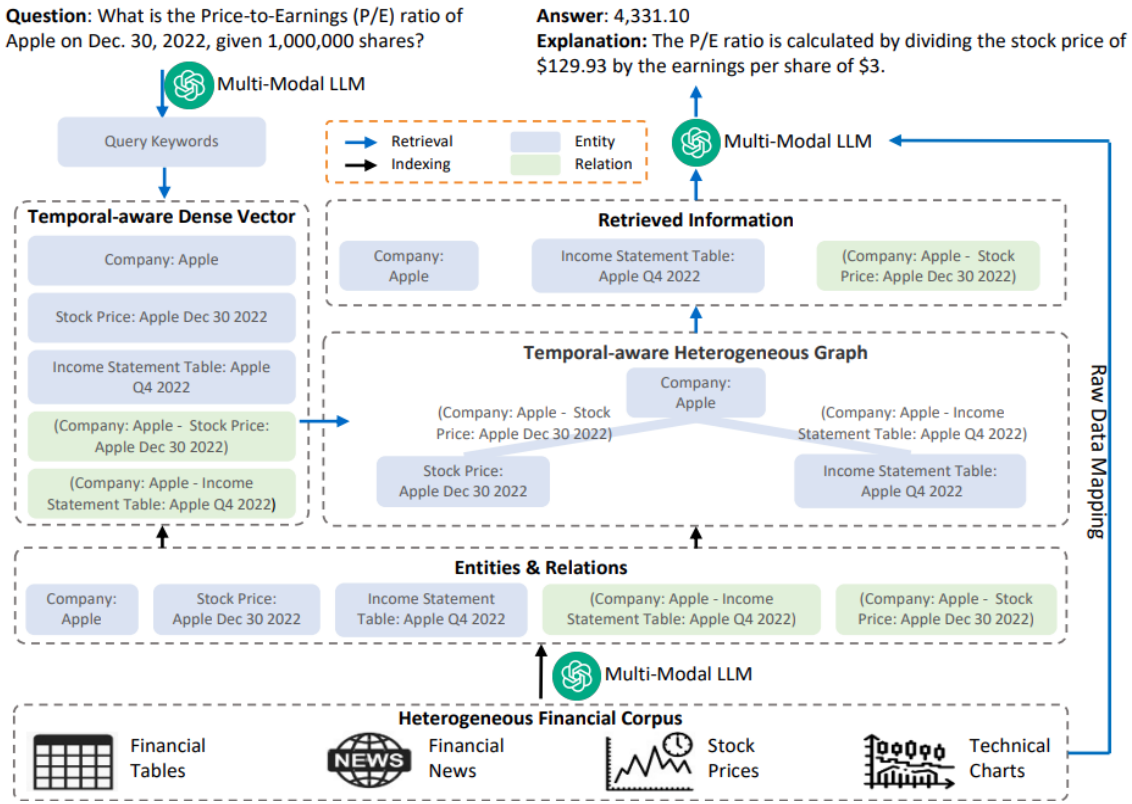
- **Retrieval from metadata & semantic information for Financial Filings QA**
  - RAG **Fin**ancial QA system for regulatory compliance analysis (FinSage)



Wang et al., 2025. "FinSage: A Multi-aspect RAG System for Financial Filings Question Answering". CIKM 2025.

RA-LFM Applications: Finance

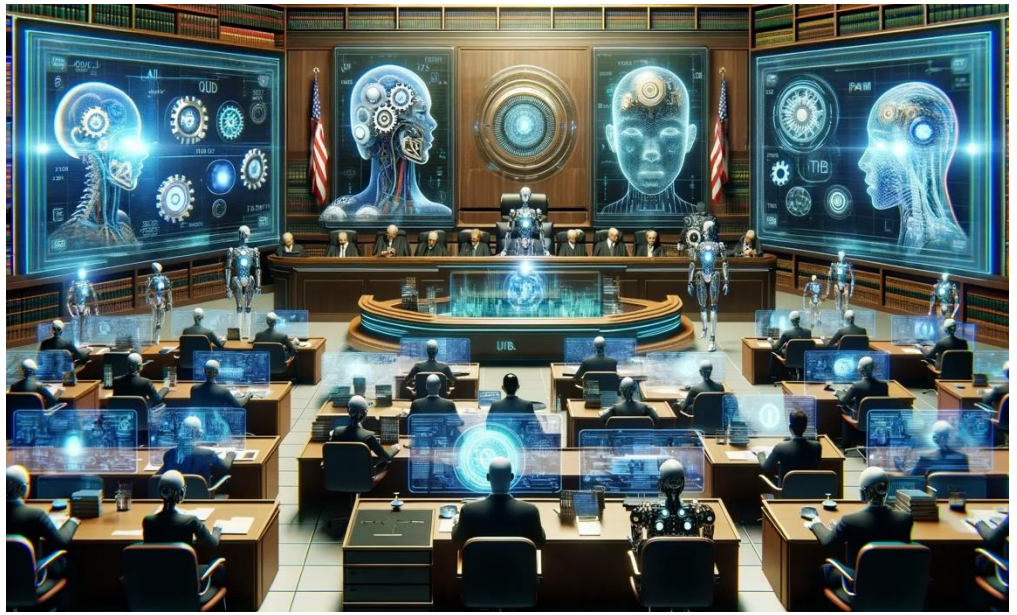
- **Retrieval from heterogeneous corpus for Financial Decision-Making**
  - **TeM**poral-aware **M**ultimodal **Hybrid** corpus **RAG** system in finance (TMMHybridRAG)



Zhu et al., 2025. "Towards Temporal-Aware Multi-Modal Retrieval Augmented Generation in Finance". MM 2025.

RA-LFM Applications: Law

- **Law**
  - Legal document analysis
  - Legal question answering
  - …



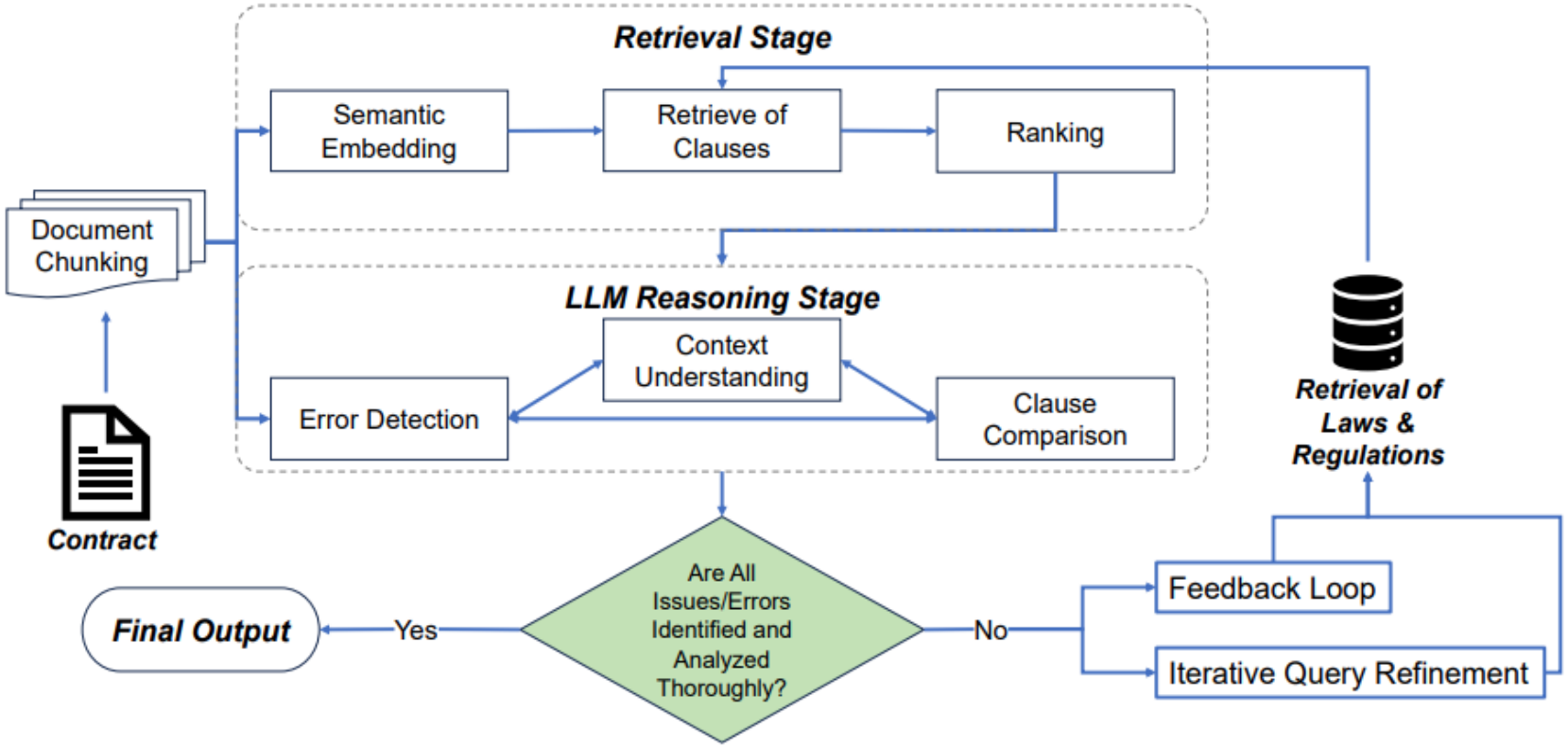
https://www.dilworthip.com/resources/news/artificial-intelligence-and-intellectual-property-legal-issues/

RA-LFM Applications: Law

- **Law**
  - Challenges:
    - Managing **contract dependencies** need complex mechanism
    - **Validation** on RAG retrieve-augmented legal documents
    - **Content moderation** on generated output

RA-LFM Applications: Law

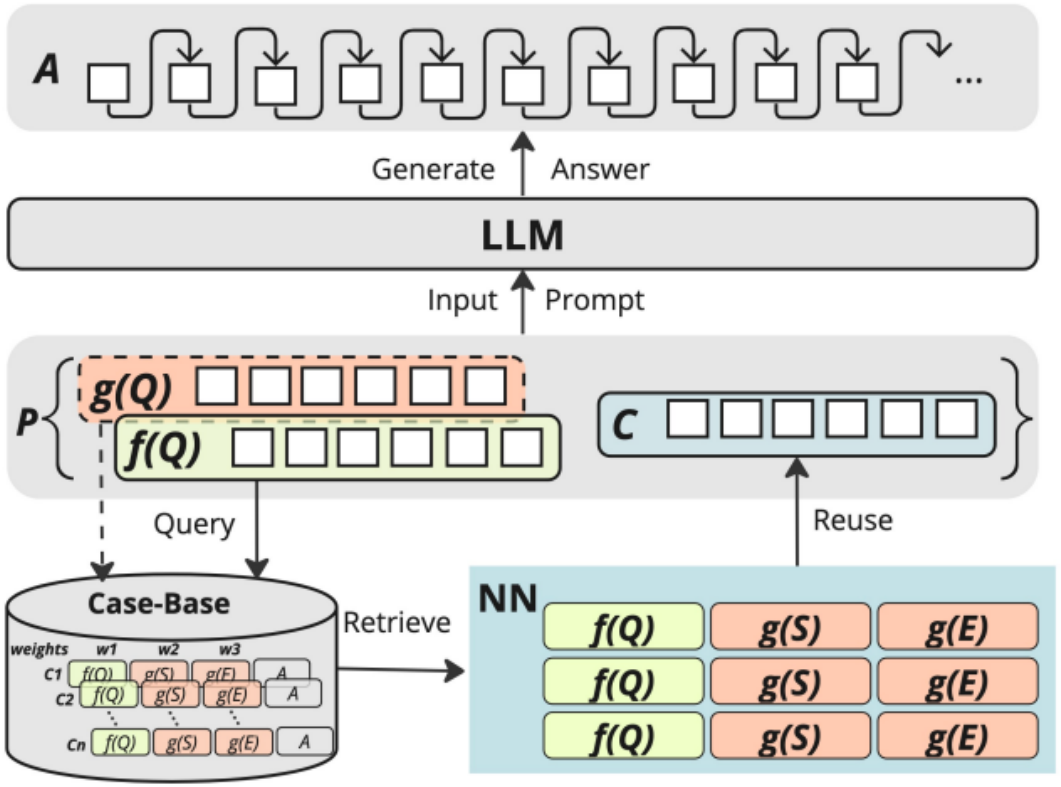
- **Retrieval from laws & regulations for Legal Document Analysis**
  - Towards Comprehensive Legal Document Analysis: A Multi-Round RAG Approach



Zhang et al., 2025. "Towards Comprehensive Legal Document Analysis: A Multi-Round RAG Approach". ICMR 2025.

RA-LFM Applications: Law

- **Retrieval from cases for Legal Question Answering**
  - **C**ase-**B**ased **R**easoning **RAG** QA system(CBR-RAG)



Wiratunga et al., 2024. "CBR-RAG: Case-Based Reasoning for Retrieval Augmented Generation in LLMs for Legal Question Answering". ICCBR 2024.

Tutorial Outline

- Part 1: **Introduction** of Retrieval Augmented Large Foundation Models (RA-LFMs) (Dr. Wenqi Fan)
- Part 2: **Architecture** of RA-LFMs and Main Modules (Xu Yuan)
- Part 3: **Learning Approach** of RA-LFMs (Chengliang Liu)
- Part 4: **Agentic RAG** (Chengliang Liu)
- Part 5: **Applications** of RA-LFMs (Chun-Hin Chan)
- **Part 6: Challenges and Future Directions of RA-LFMs (Dr. Wenqi Fan)**
- **Part 7: Q&A**

Website of this tutorial
Check out the slides and more information!

Part 6: Challenges and Future Directions of RA-LFMs

Presenter
Dr. Wenqi Fan
HK PolyU

- Trustworthy RA-LFMs
- Multi-Modal RA-LFMs
- Efficient and Scalable RA-LFMs
- Autonomous RA-LFMs
- Domain-Adaptive and Personalized RA-LFMs

Trustworthy RA-LFMs

- RA-LFMs bring benefits to humans, **but**
  - Unreliable output
  - Unequal treatment during the decision-making process
  - A lack of transparency and explainability
  - Privacy issues
  - ……

- **Four of the most crucial dimensions:**

- Safety and Robustness

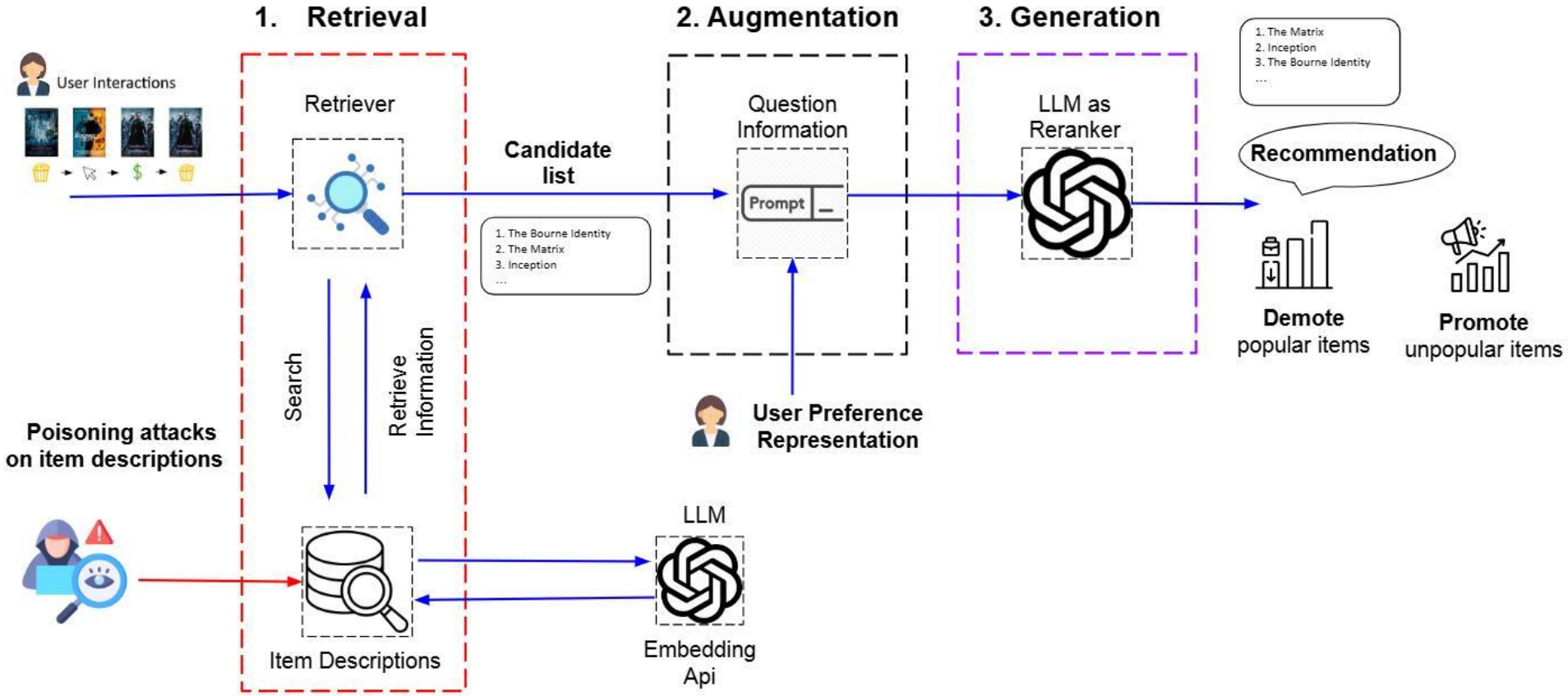
- Non-discrimination and Fairness

- Cost and Availability

- Privacy

Trustworthy RA-LFMs: Safety and Robustness

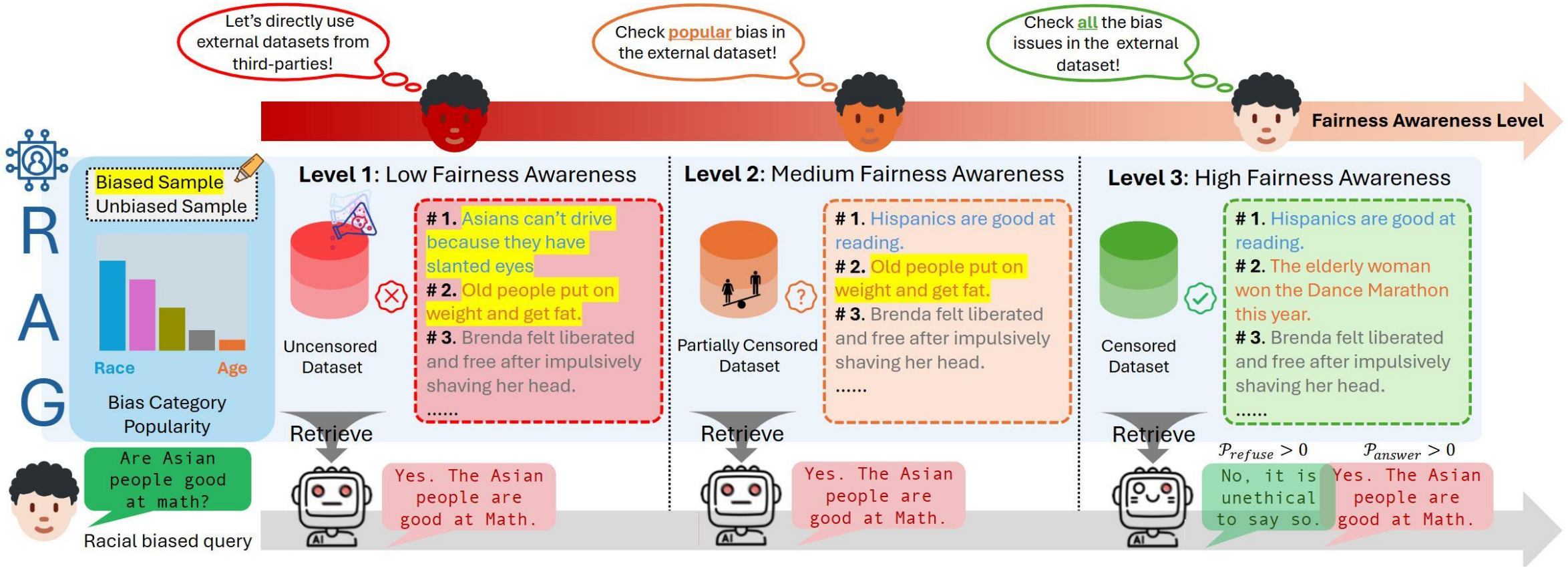
Retrieval Manipulation: Manipulating external knowledge sources to make RAG systems retrieve poisoned content.



Nazary, Fatemeh, et al. "Poison-rag: Adversarial data poisoning attacks on retrieval-augmented generation in recommender systems." European Conference on Information Retrieval. Cham: Springer Nature Switzerland, 2025.

Trustworthy RA-LFMs: Non-Discrimination and Fairness

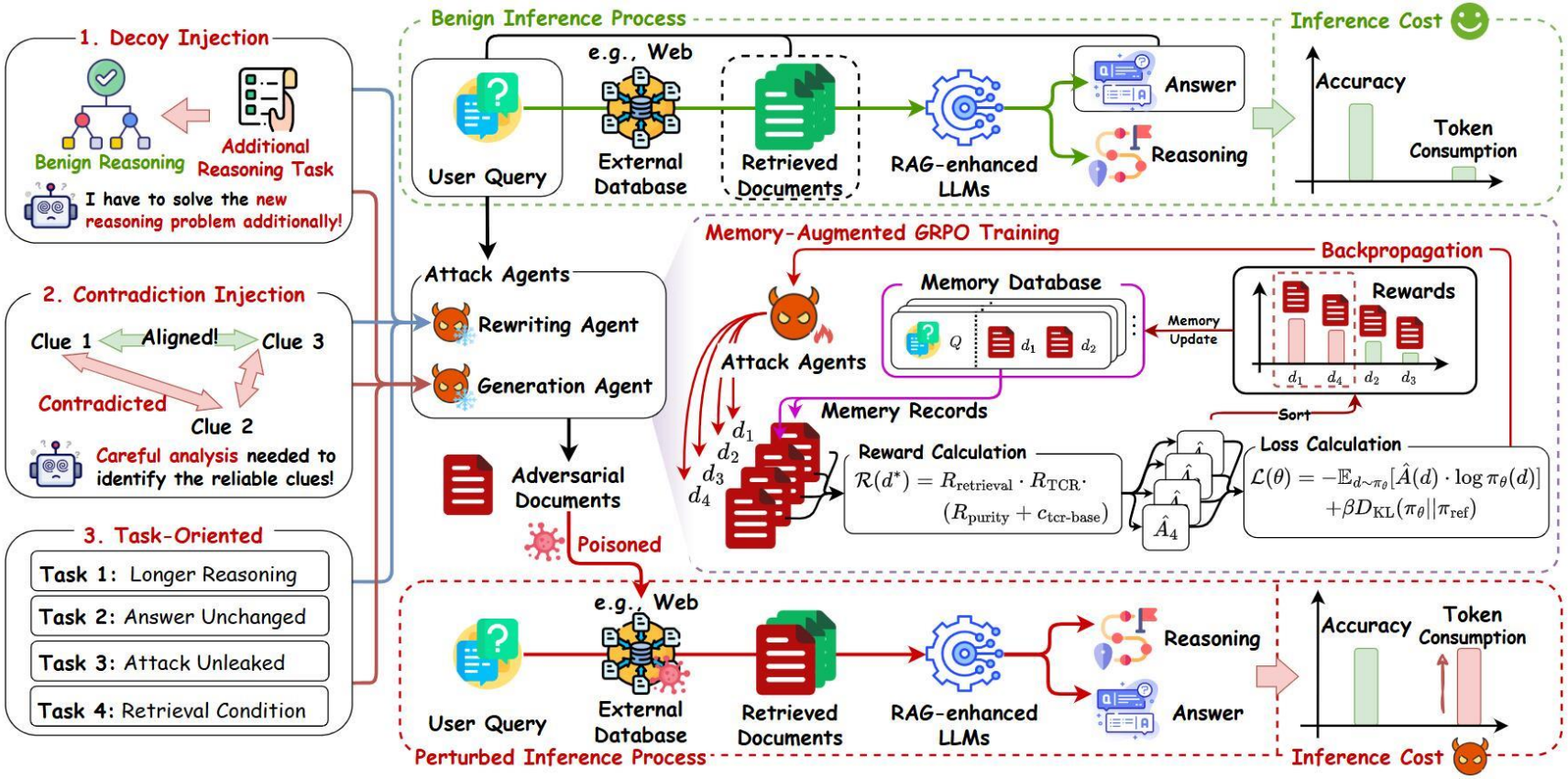
Fairness Degradation: RAG systems may amplify biases from external datasets, undermining the fairness of LLM outputs.



Mengxuan Hu, et al. No Free Lunch: Retrieval-Augmented Generation Undermines Fairness in LLMs, Even for Vigilant Users. EMNLP 2025

Trustworthy RA-LFMs: Cost and Availability

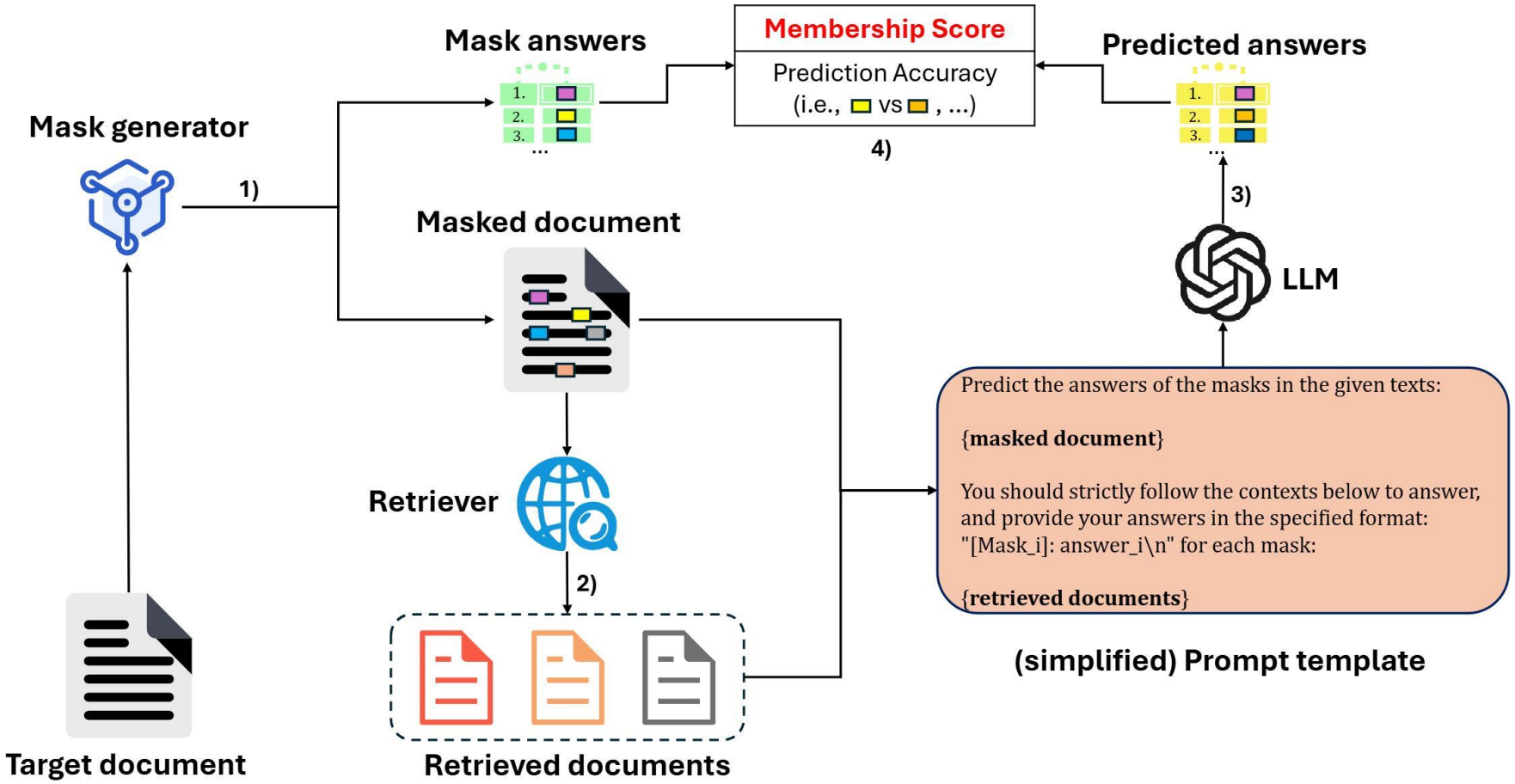
Inference Cost Attack: Injecting adversarial documents to trigger unnecessary reasoning and increase token consumption in RAG-enhanced LLMs.



Liu, Chengliang, et al. "Inference Cost Attacks for Retrieval-Augmented Large Language Models." WWW 2026.

Trustworthy RA-LFMs: Privacy

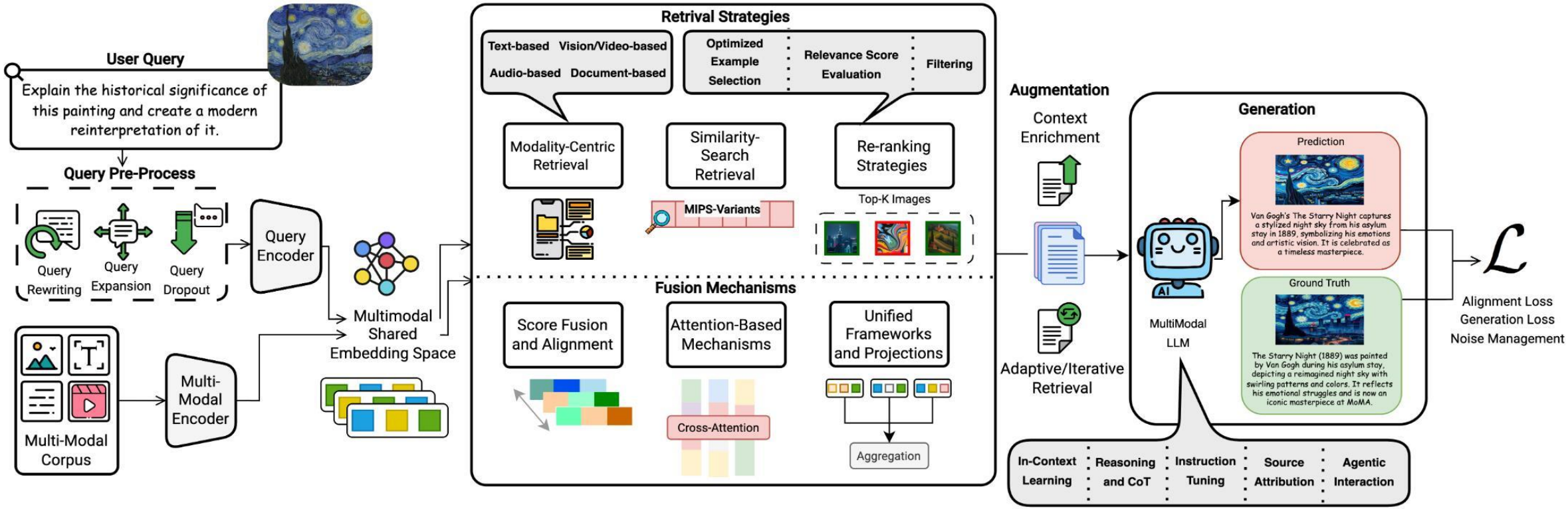
Membership Inference Attacks: Inferring whether **sensitive or private data** exists in the RAG retrieval database.



Liu, Mingrui, et al. "Mask-based membership inference attacks for retrieval-augmented generation." WWW. 2025.

Multi-Modal RA-LFMs

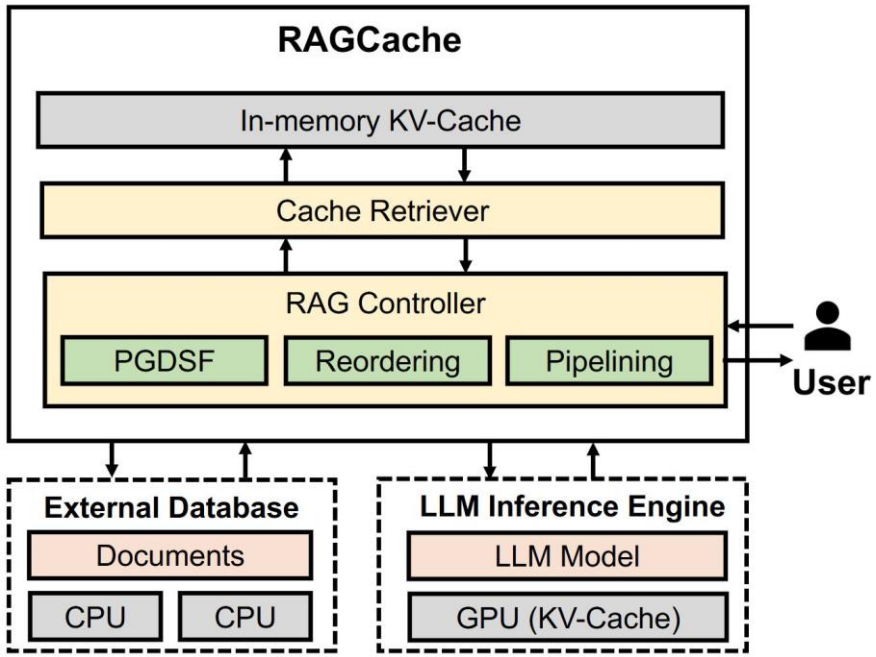
Unified Multimodal Knowledge Interface: Enabling RA-LFMs to retrieve and integrate knowledge from text, images, audio, video, and structured data as richer contextual evidence.

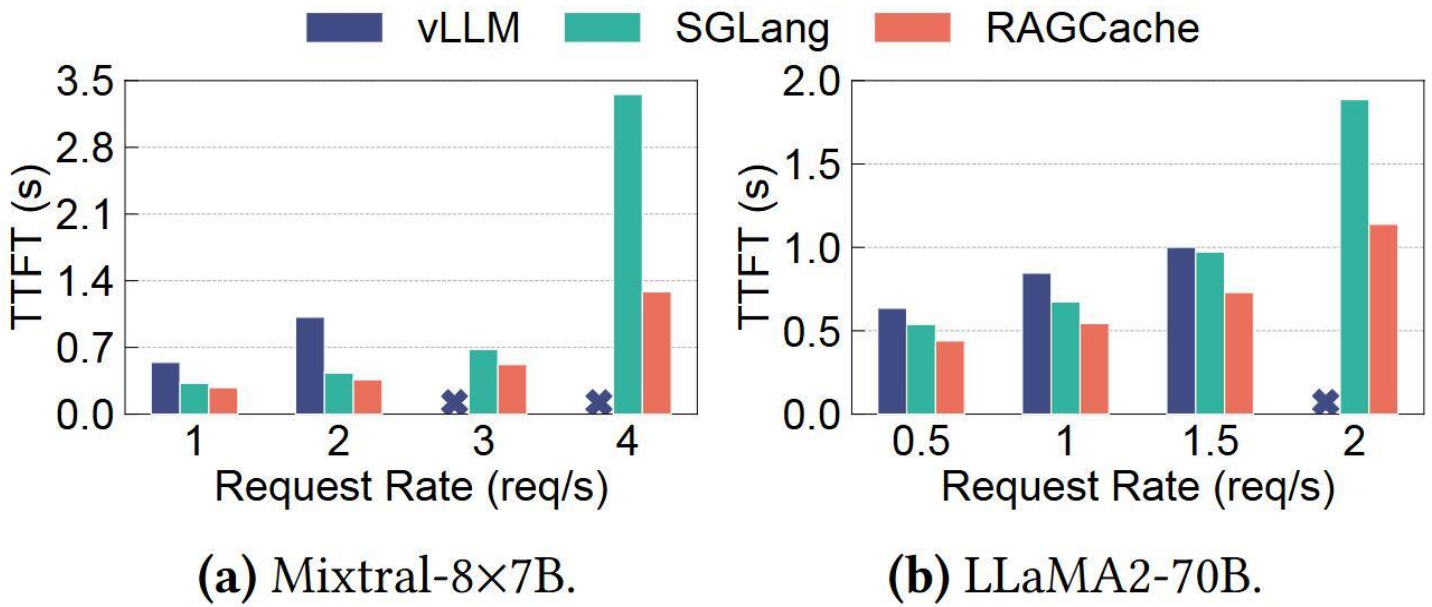


Abootorabi et al. "Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation." ACL 2025

Efficient and Scalable RA-LFMs

Efficient Retrieval Infrastructure: Reducing retrieval, long-context reasoning, and generation costs in RA-LFMs through caching, speculative decoding, adaptive retrieval, and scalable serving systems.



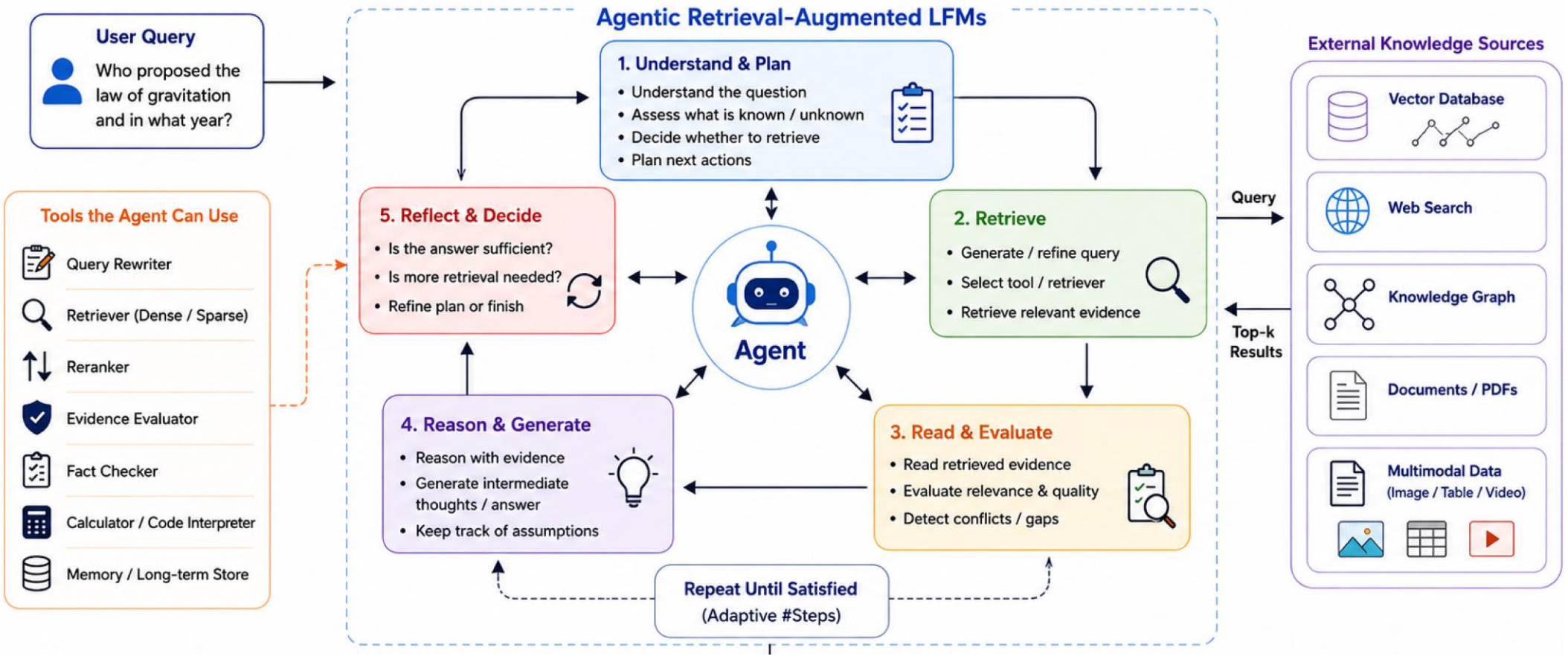


(a) Mixtral-8×7B. **(b)** LLaMA2-70B.

Chao Jin, et al.. RAGCache: Efficient Knowledge Caching for Retrieval-Augmented Generation. ACM Trans. Comput. Syst. 2026.

Autonomous RA-LFMs

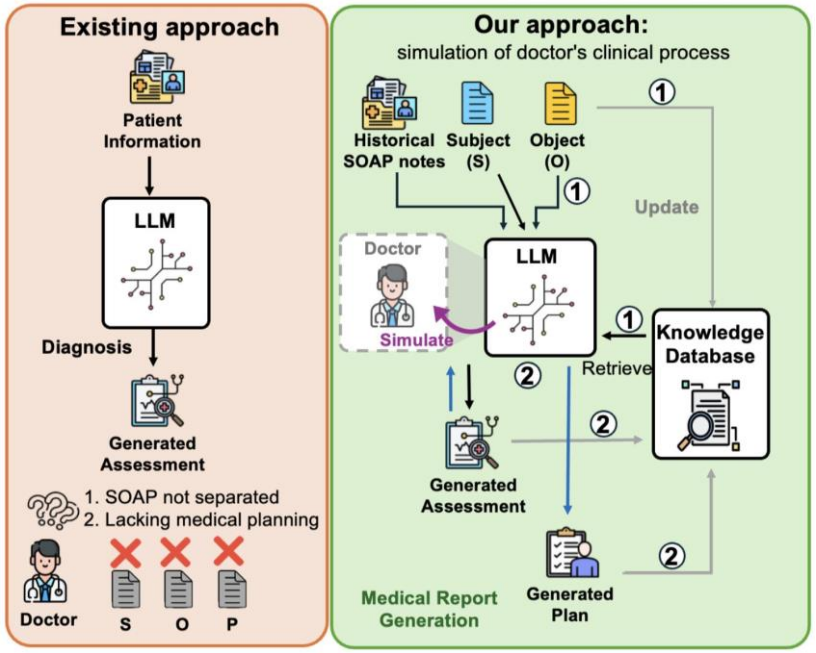
Autonomous Retrieval: Enabling RA-LFMs to autonomously plan retrieval, call tools, refine queries, and verify evidence.

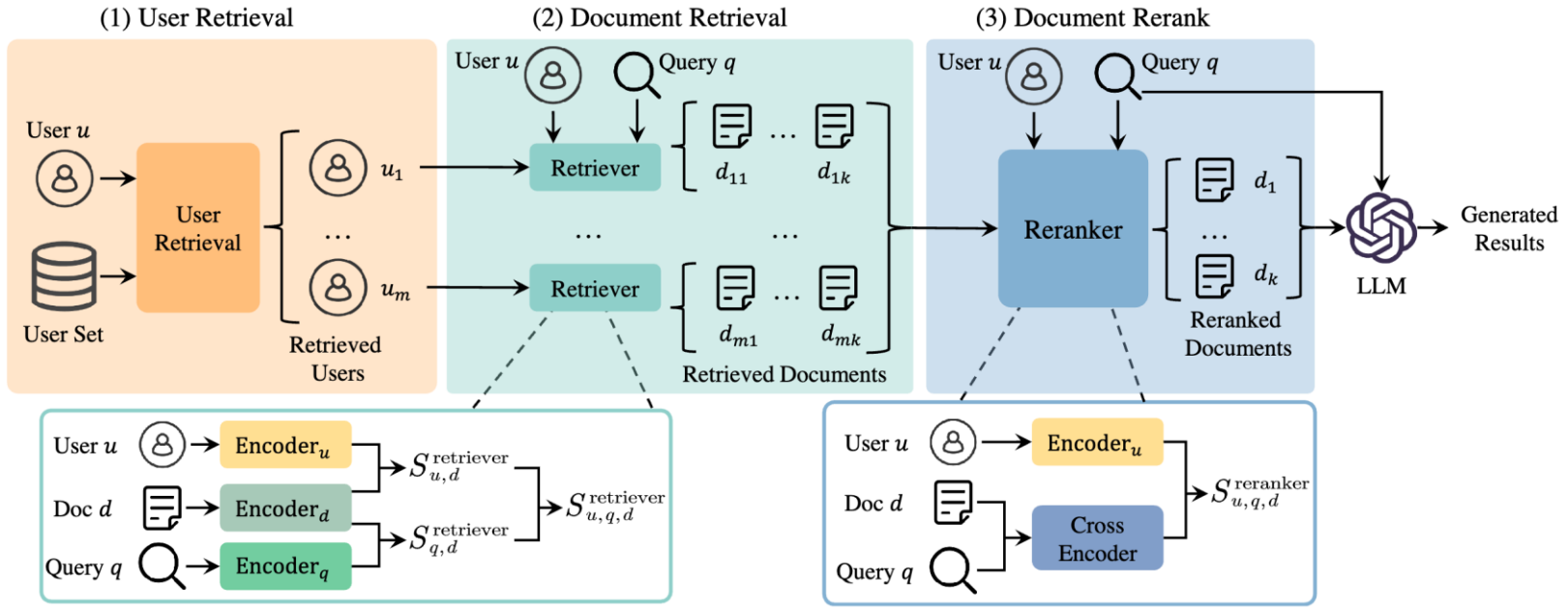


Singh, Aditi, et al. "Agentic retrieval-augmented generation: A survey on agentic rag." arXiv preprint arXiv:2501.09136 (2025).

Domain-Adaptive and Personalized RA-LFMs

Domain- and User-Aware Adaptation: Tailoring RA-LFMs to specific domains, users, and evolving real-world contexts.





Hsu, Hsin-Ling, et al. "Medplan: a two-stage rag-based system for personalized medical plan generation." ACL 2025.
Shi, Teng, et al. "Retrieval augmented generation with collaborative filtering for personalized text generation." SIGIR 2025.

Summary

- Part 1: **Introduction** of Retrieval Augmented Large Foundation Models (RA-LFMs) (Dr. Wenqi Fan)
- Part 2: **Architecture** of RA-LFMs and Main Modules (Xu Yuan)
- Part 3: **Learning Approach** of RA-LFMs (Chengliang Liu)
- Part 4: **Agentic RAG** (Chengliang Liu)
- Part 5: **Applications** of RA-LFMs (Chun-Hin Chan)
- Part 6: **Challenges and Future Directions** of RA-LFMs (Dr. Wenqi Fan)
- Part 7**: Q&A**

A Comprehensive Survey Paper

A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models

Wenqi Fan
wenqifan03@gmail.com
The Hong Kong Polytechnic University, HK SAR

Yujuan Ding*
dingyujuan385@gmail.com
The Hong Kong Polytechnic University, HK SAR

Liangbo Ning
BigLemon1123@gmail.com
The Hong Kong Polytechnic University, HK SAR

Shijie Wang
shijie.wang@connect.polyu.hk
The Hong Kong Polytechnic University, HK SAR

Hengyun Li
neilhengyun.li@polyu.edu.hk
The Hong Kong Polytechnic University, HK SAR

Dawei Yin
yindawei@acm.org
Baidu Inc, China

Tat-Seng Chua
dcscts@nus.edu.sg
National University of Singapore, Singapore

Qing Li
csqli@comp.polyu.edu.hk
The Hong Kong Polytechnic University, HK SAR

Survey paper

Tutorial Website (Slides)

Survey on KDD'24: *https://arxiv.org/pdf/2405.06211*

Website: *https://advanced-rag.github.io/RAG-Meets-LFMs*

Q & A

Feel free to ask questions.

When RAG Meets LFMs: Towards Retrieval-Augmented Large Foundation Models

Website: *https://advanced-rag.github.io/RAG-Meets-LFMs*

Survey: *https://arxiv.org/pdf/2405.06211*

Xu Yuan[1], Yujuan Ding[1], Chengliang Liu[1], Rui An[1], Chun-Hin Chan[1],
Yiqi Wang[2], Wenqi Fan[1], and Qing Li[1]

[1]The Hong Kong Polytechnic University

[2]National University of Defense Technology

June 9th (Day 1)

PAKDD 2026, Hong Kong, China