Tutorial Outline

- **Part 1: Introduction** of Retrieval Augmented Large Language Models (RA-LLMs) (Dr. Wenqi Fan)
- **Part 2: Architecture** of RA-LLMs and **Main Modules** (Dr. Yujuan Ding)
- **Part 3: Learning Approach of RA-LLMs (Liangbo Ning)**
- **Part 4: Applications** of RA-LLMs (Shijie Wang)
- **Part 5: Challenges and Future Directions** of RA-LLMs (Dr. Wenqi Fan)

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

Website

Part 3: RA-LLM Learning

Presenter
Liangbo Ning
HK PolyU

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

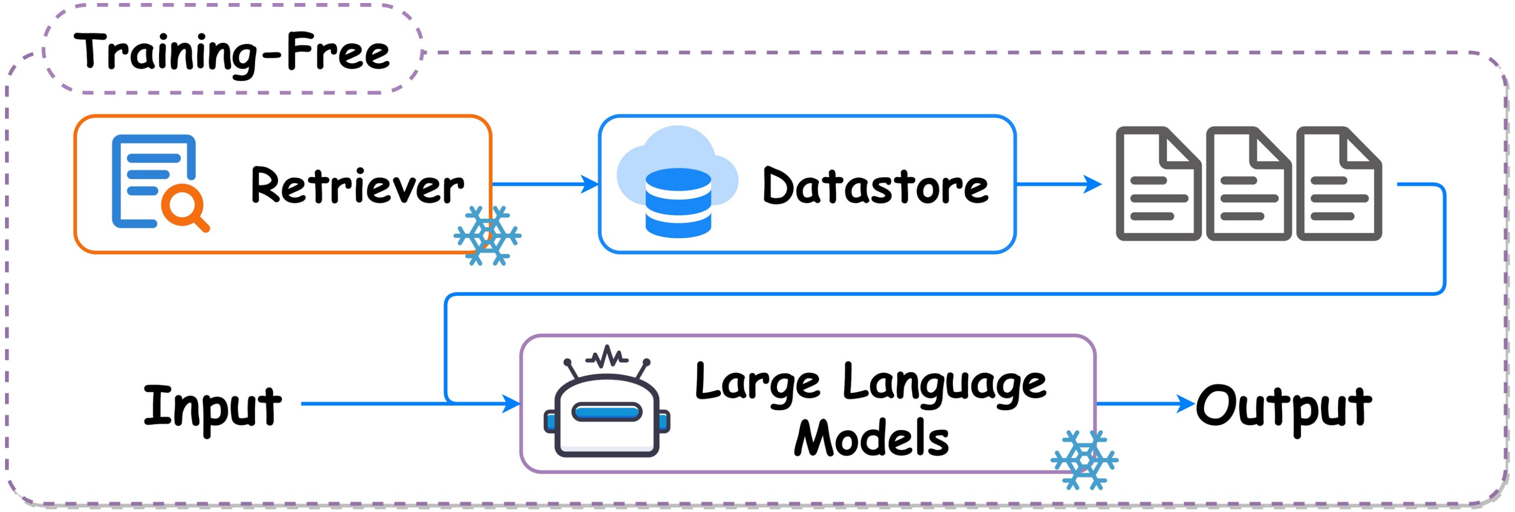
Part 3: RA-LLM Learning

Presenter
Liangbo Ning
HK PolyU

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

RA-LLM Learning: Training-free

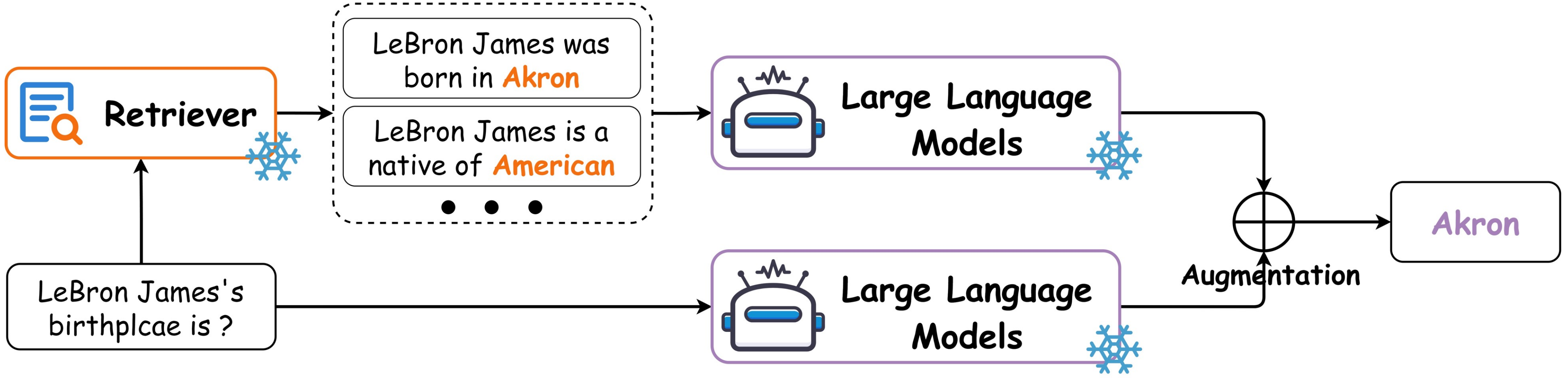
- **Retrieval models** and **language models** are both **frozen**.



RA-LLM Learning: Training-free

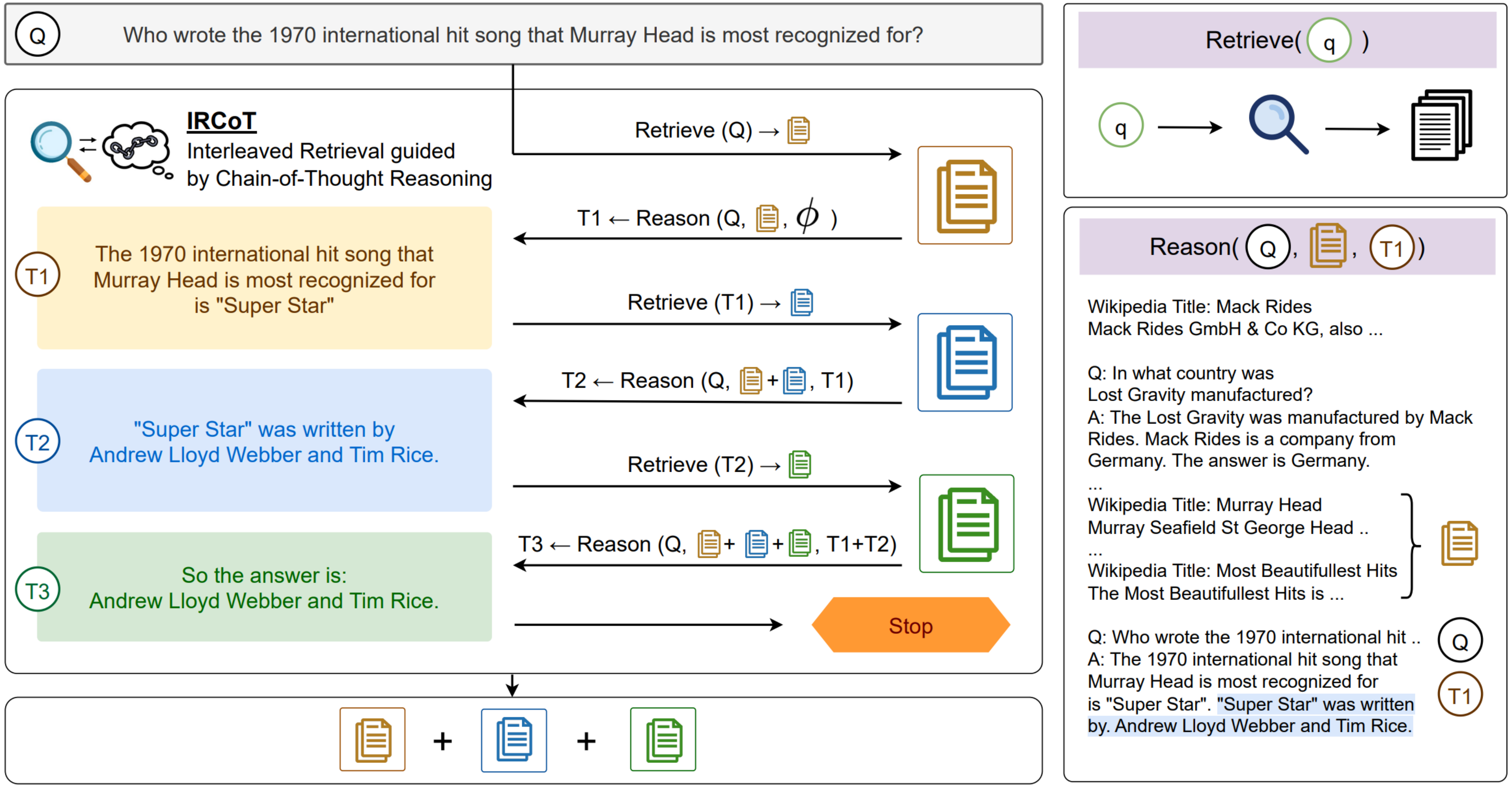
- **Prompt Engineering-based Methods**

- **Retrieval-Guided Token Generation Methods**



RA-LLM Learning: Training-free

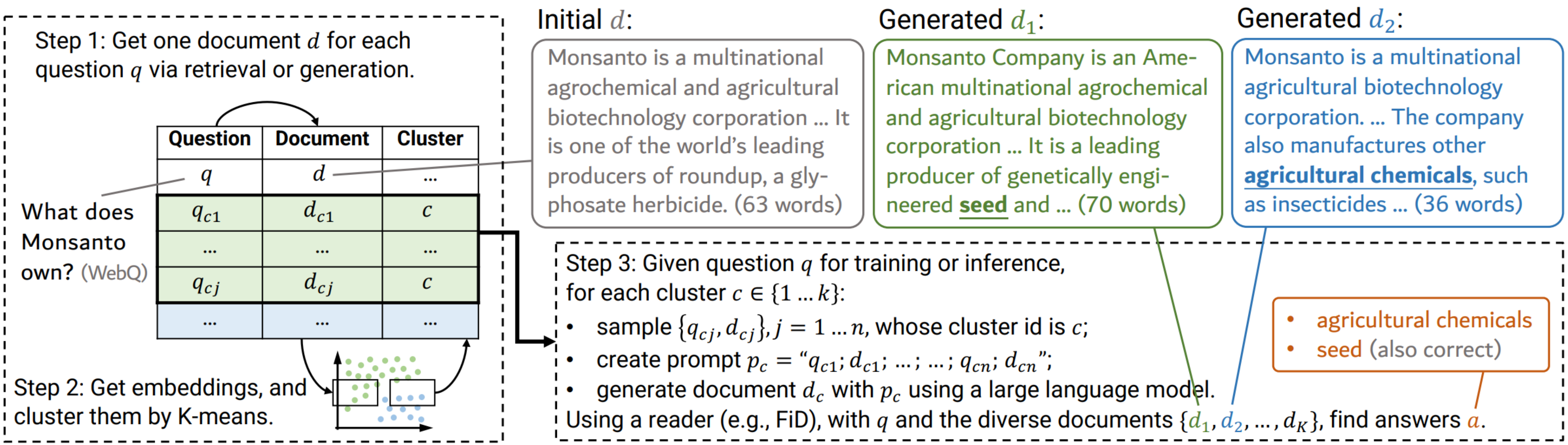
- **IRCoT**



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

RA-LLM Learning: Training-free

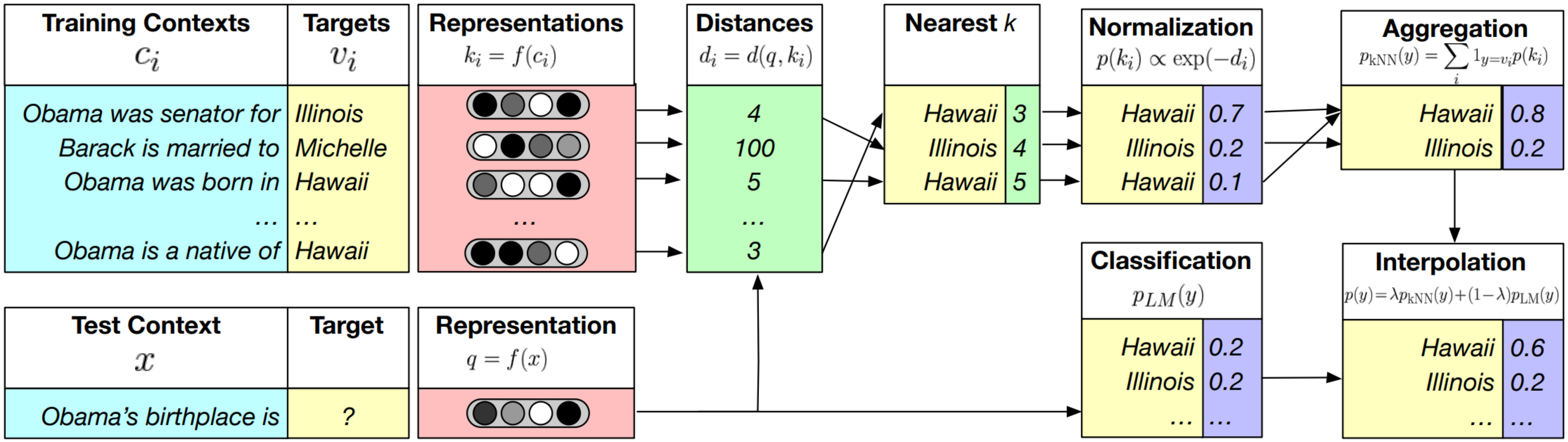
- **GENREAD**



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

RA-LLM Learning: Training-free

- ***k*NN-LM**

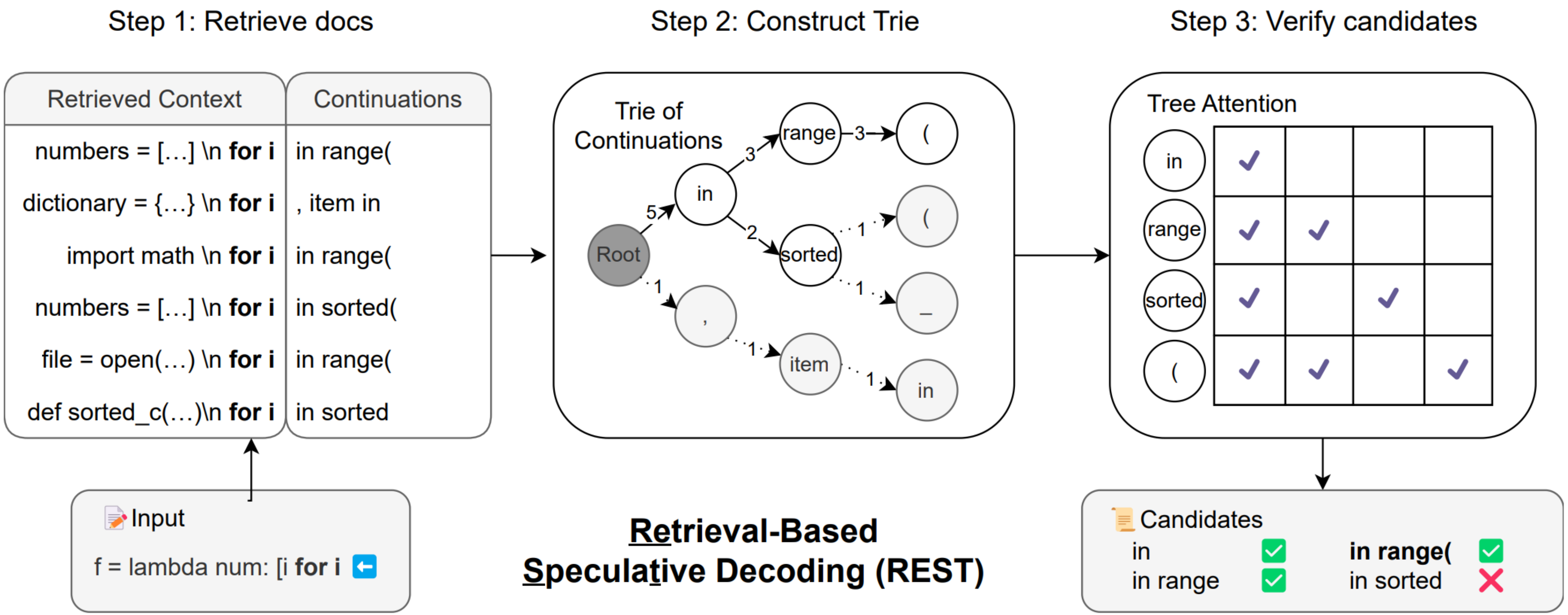


$$p(y|x) = \lambda\, p_{\mathrm{kNN}}(y|x) + (1-\lambda)\, p_{\mathrm{LM}}(y|x)$$

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

RA-LLM Learning: Training-free

- **REST**



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

RA-LLM 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-LLM Learning

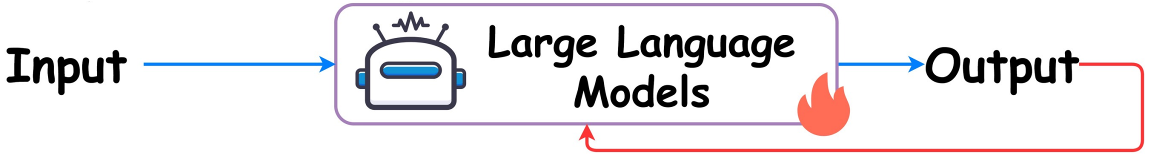
Website

Website of this tutorial

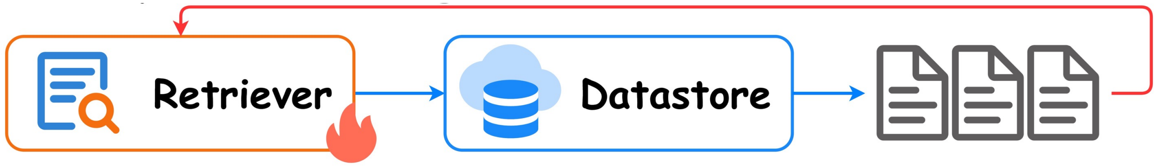
- Training-free Methods
- **Training-based Methods**
  - Independent Learning
  - Sequential Learning
  - Joint Learning

RA-LLM Learning: Independent Training

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

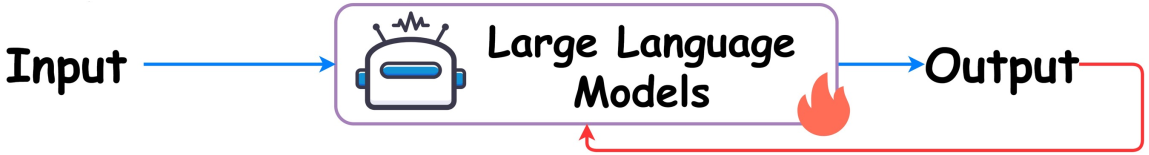


    - **Independent training of Retriever.**

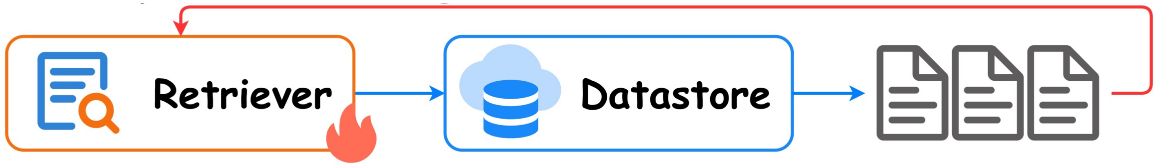


RA-LLM Learning: Independent Training

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

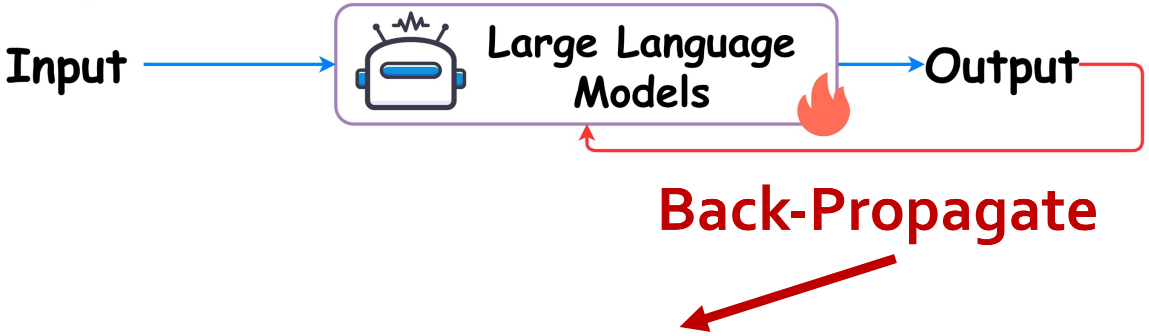


  - Independent training of Retriever.



RA-LLM Learning: Independent Training

- Independent training of large language models.

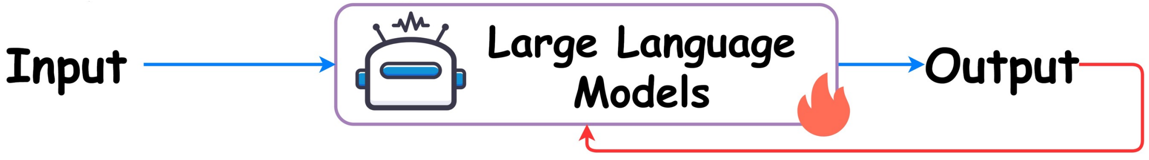


$$Minimize \ -logP_{LM}(y|x)$$

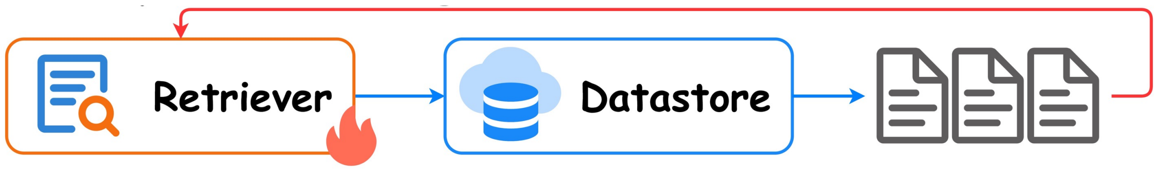
OpenAI | Meta | Google AI | T5 | ......

RA-LLM Learning: Independent Training

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



  - **Independent training of Retriever.**



RA-LLM Learning: Independent Training

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

Kobe Bryant, a legendary basketball player, left an indelible mark on the sport …

Kobe Bryant, a basketball icon and five-time NBA champion, captivated fans worldwide …

Text Chunks

[0, 0.8, 0, 0.9, 0.7, …]

[1.2, 0, 0, 0.6, 0.8, …]

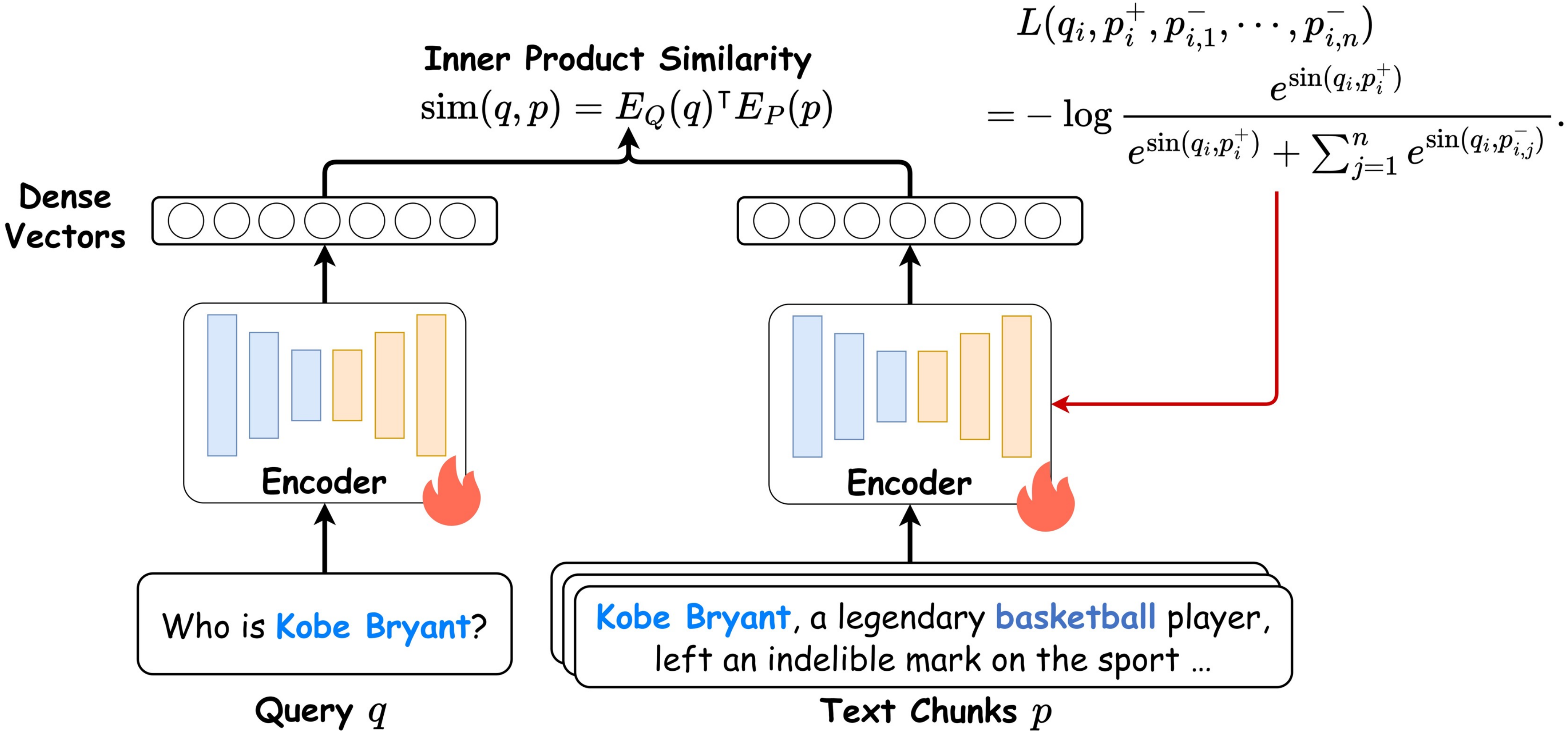
Sparse Vectors

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-LLM Learning: Independent Training

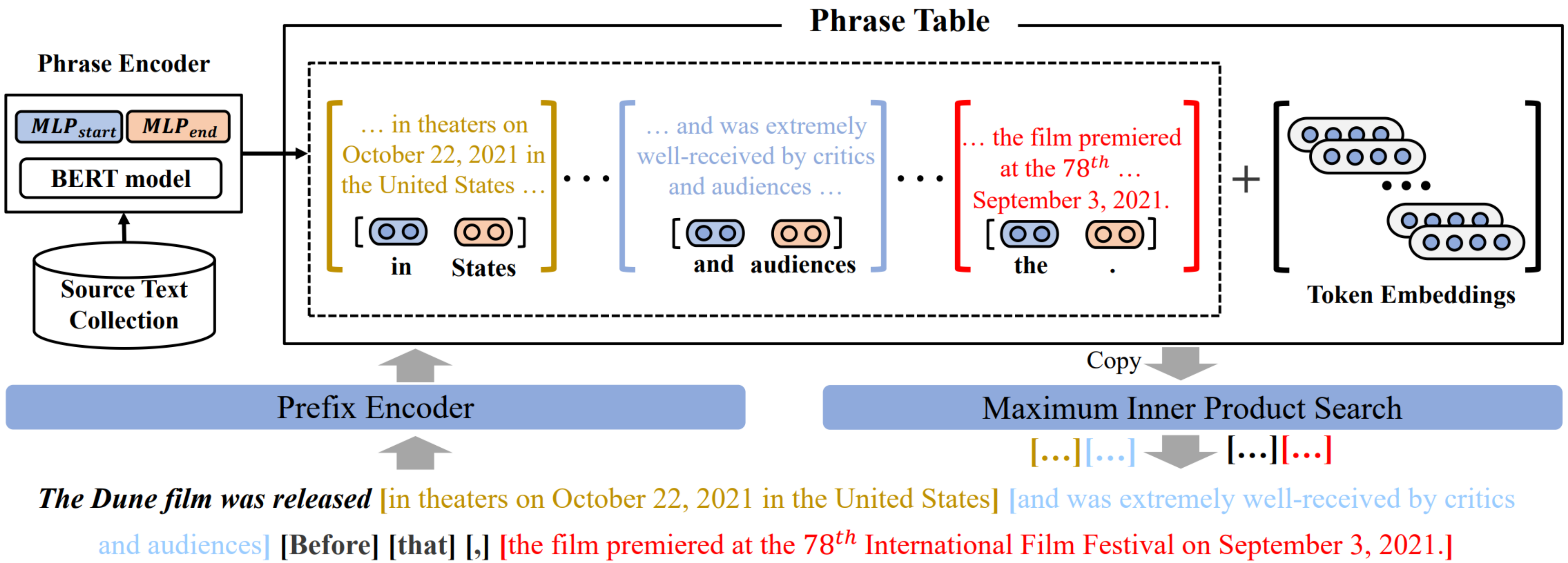
- **Dense retrieval models: DPR**



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

RA-LLM Learning: Independent Training

- **Dense retrieval models: CoG**



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

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

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

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

RA-LLM 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-LLM 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

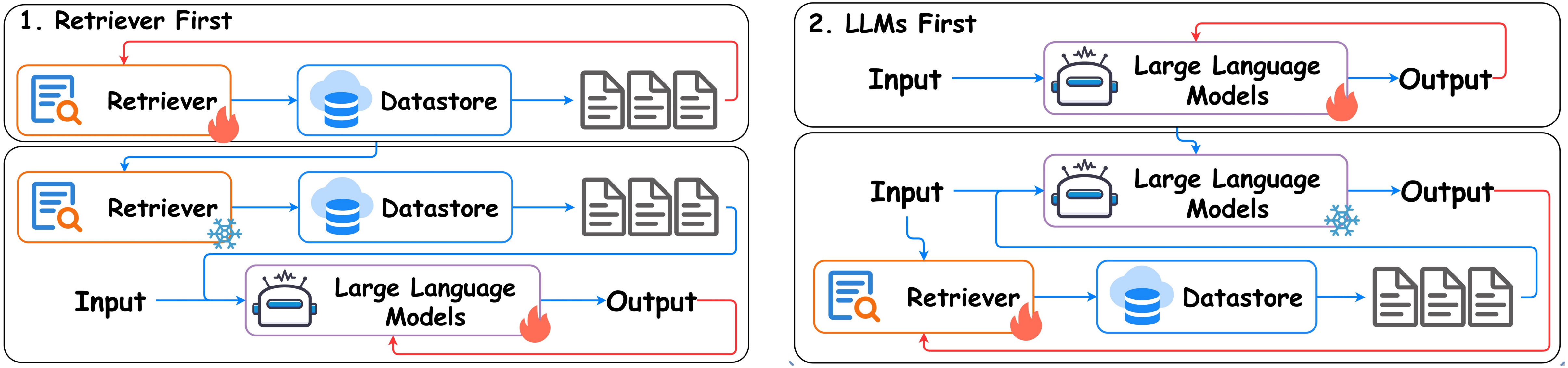
Part 3: RA-LLM Learning

Website of this tutorial

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

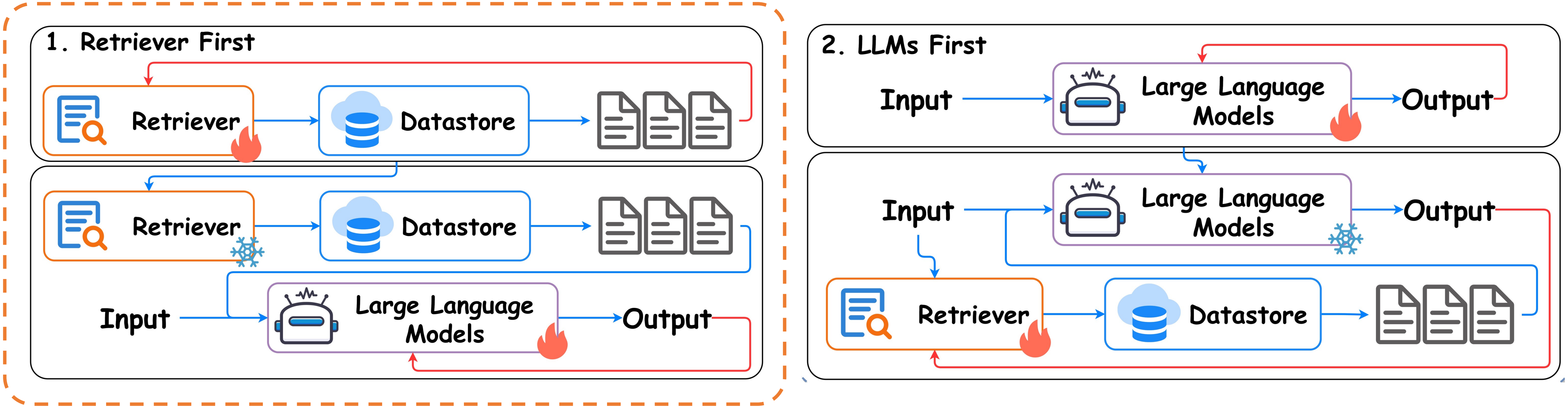
RA-LLM Learning: Sequential Training

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



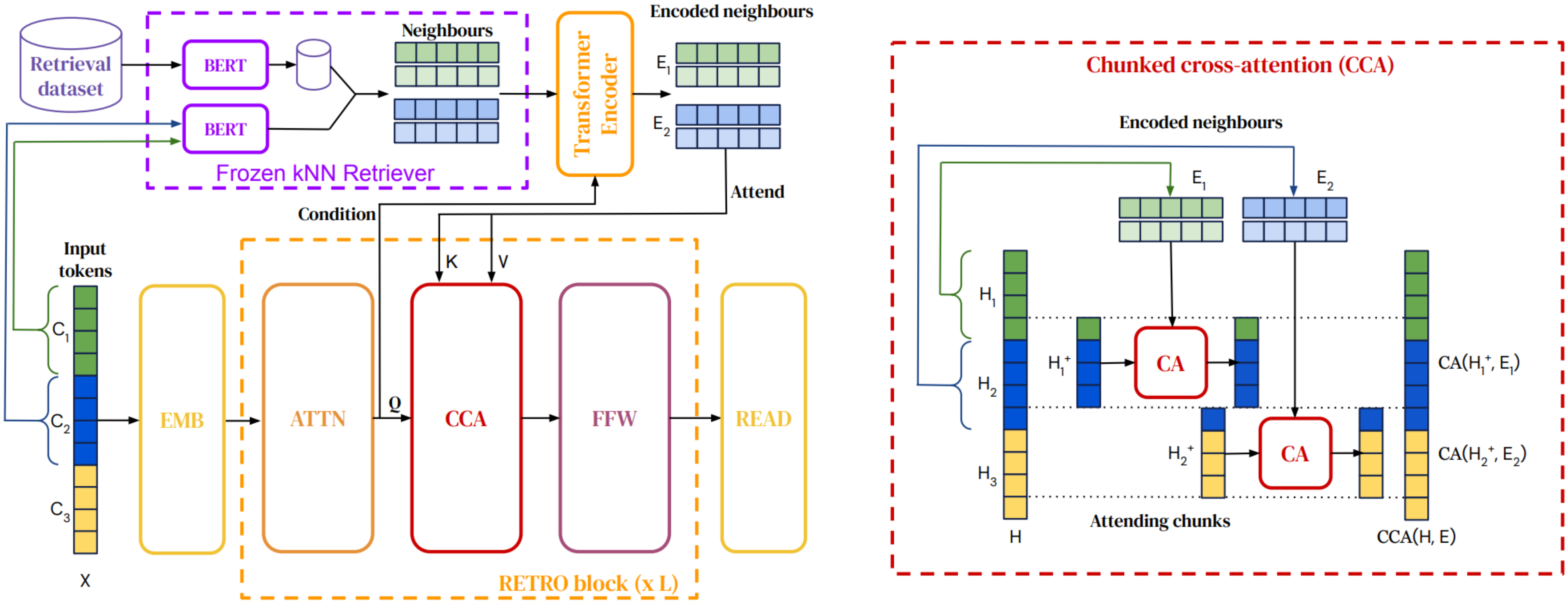
RA-LLM Learning: Sequential Training

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



RA-LLM Learning: Sequential Training

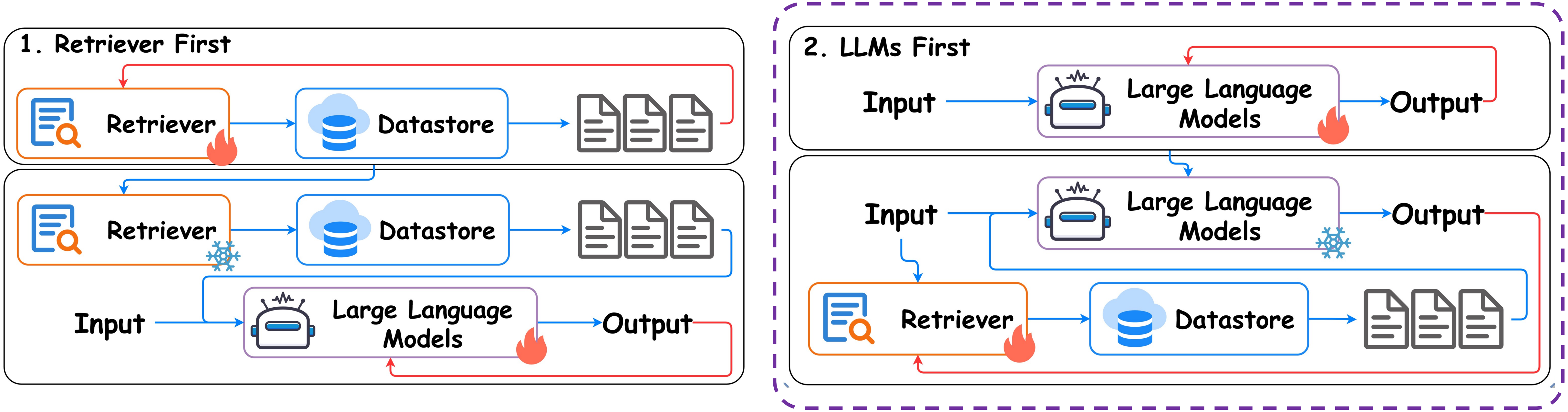
- **RETRO**



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

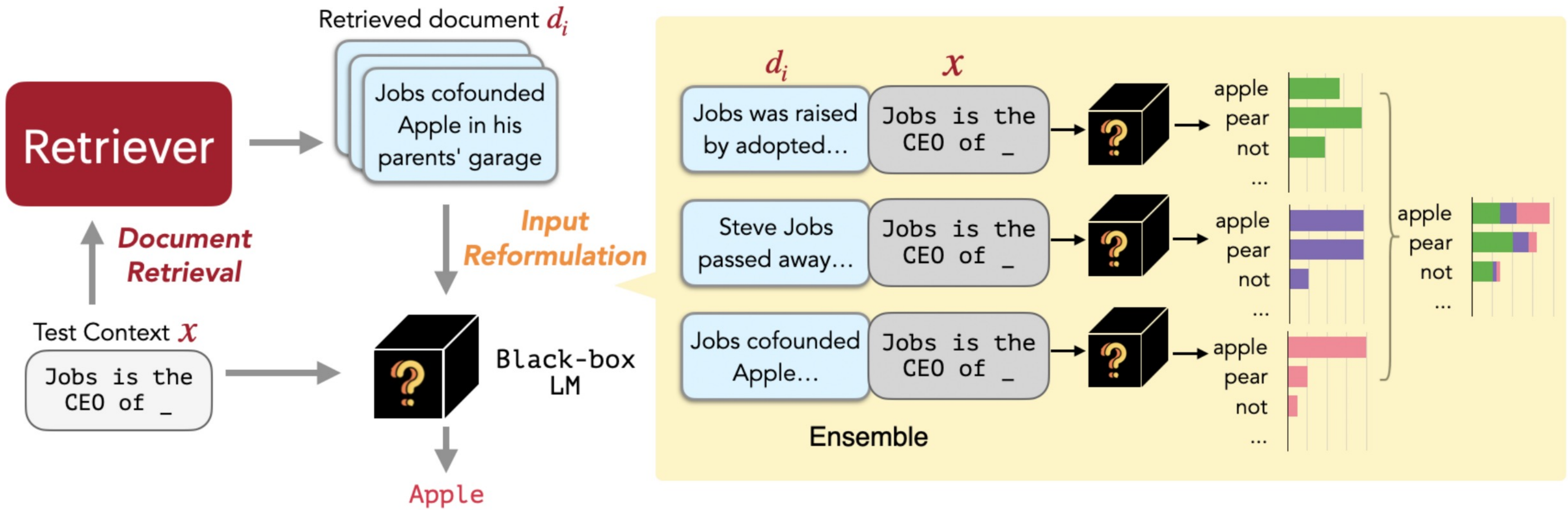
RA-LLM Learning: Sequential Training

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



RA-LLM Learning: Sequential Training

- **REPLUG (Retrieve and Plug)**

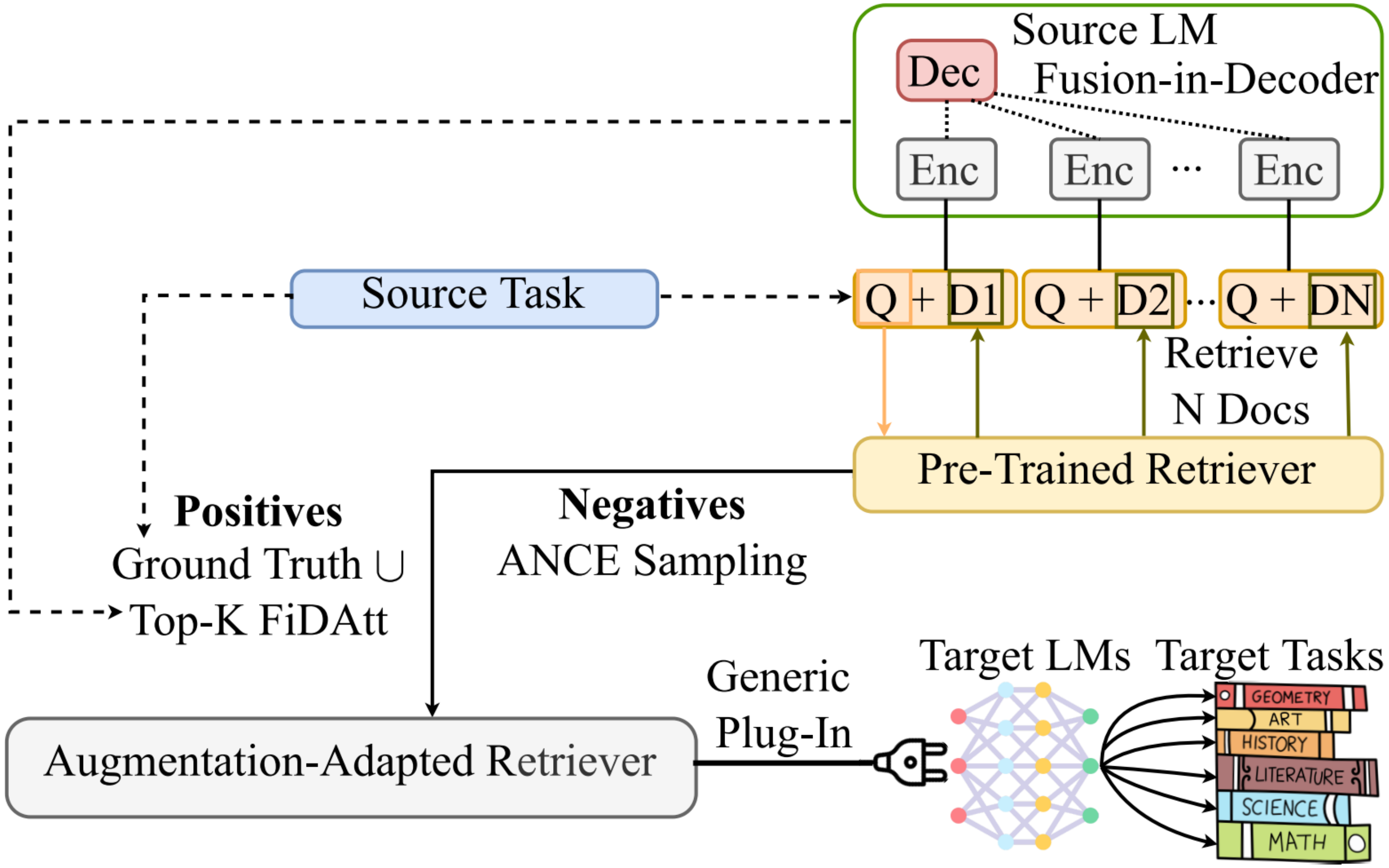


$$\mathcal{L} = \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}} KL\Big(P_R(d \mid x) \parallel Q_{\text{LM}}(d \mid x, y)\Big) \qquad P_R(d \mid x) = \frac{e^{s(d,x)/\gamma}}{\sum_{d\in\mathcal{D}'} e^{s(d,x)/\gamma}} \qquad Q(d \mid x, y) = \frac{e^{P_{LM}(y|d,x)/\beta}}{\sum_{d\in\mathcal{D}'} e^{P_{LM}(y|d,x)/\beta}}$$

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

RA-LLM 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-LLM 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 modell

Part 3: RA-LLM Learning

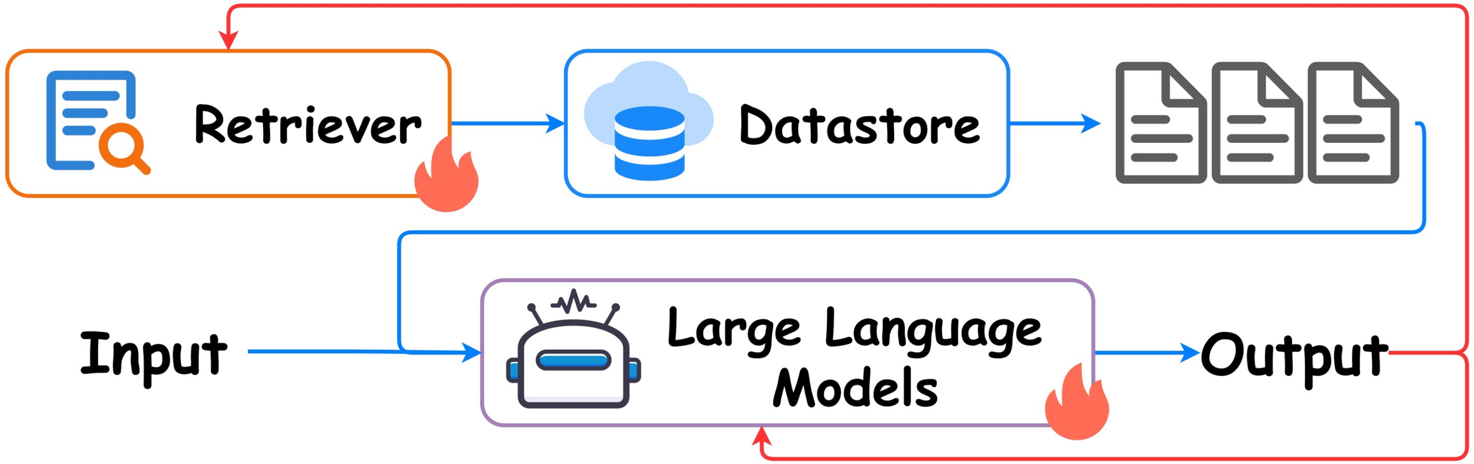
Website

Website of this tutorial

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

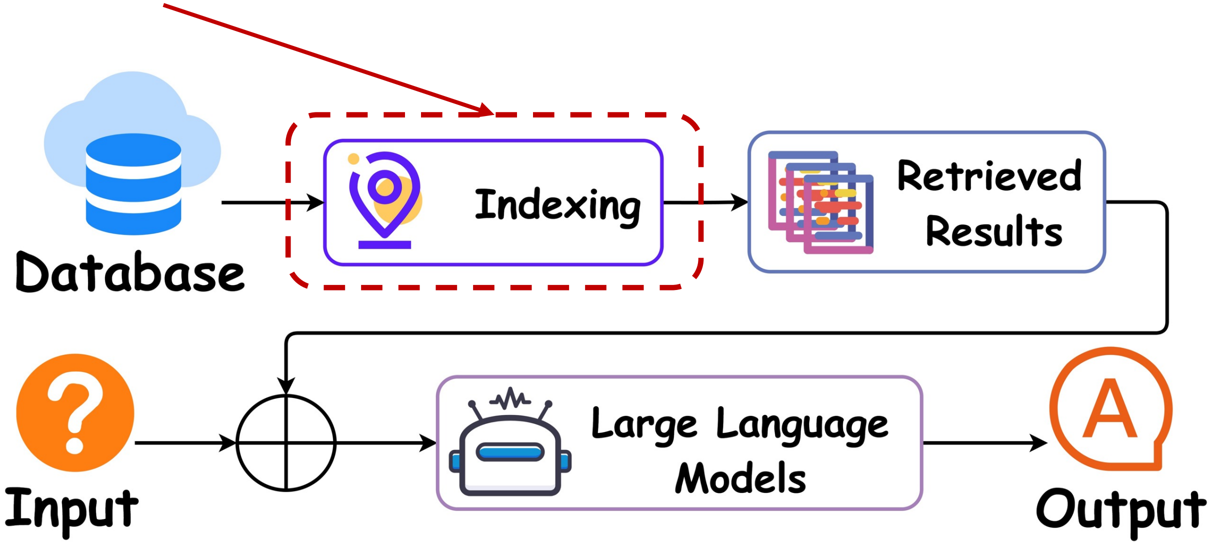
RA-LLM Learning: Joint Training

- Retrieval models is and language models are trained jointly.



RA-LLM Learning: Joint Training

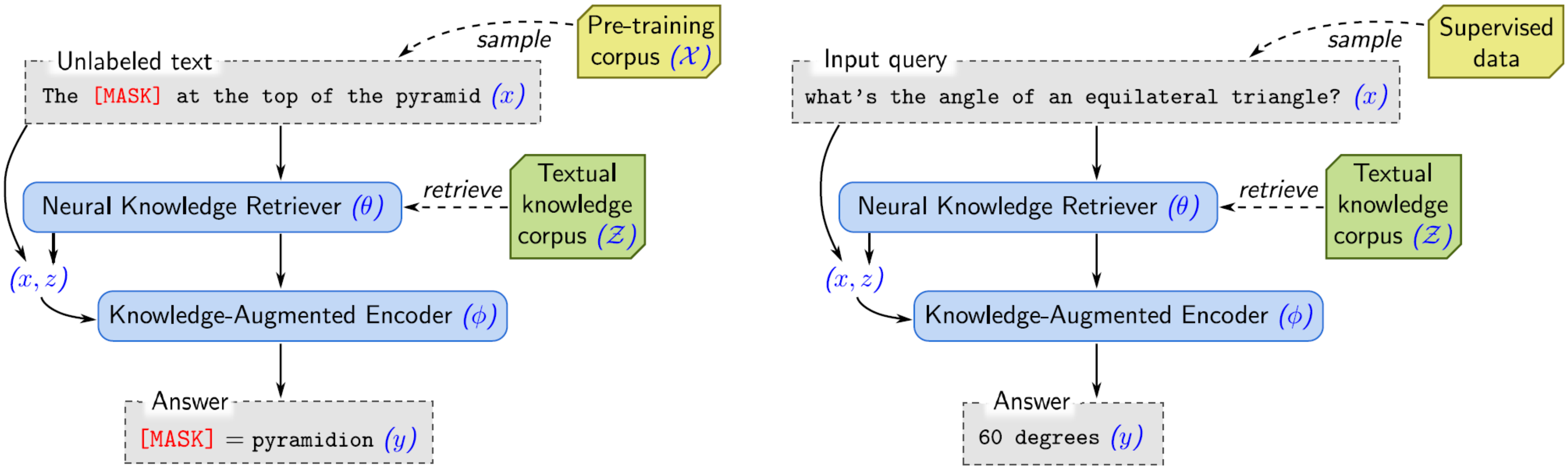
- **Retrieval Index Updating, which could be very expensive!**



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

RA-LLM Learning: Joint Training

- **REALM**

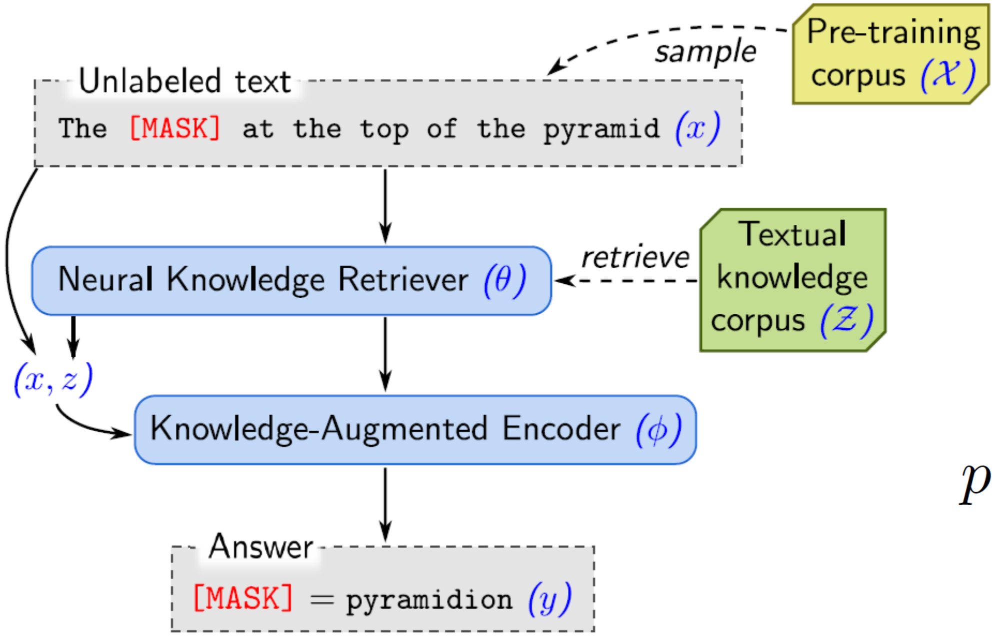


Objective function: $p(y \mid x) = \sum_{z \in \mathcal{Z}} p(y \mid z, x)\, p(z \mid x).$

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

RA-LLM Learning: Joint Training

- **REALM**



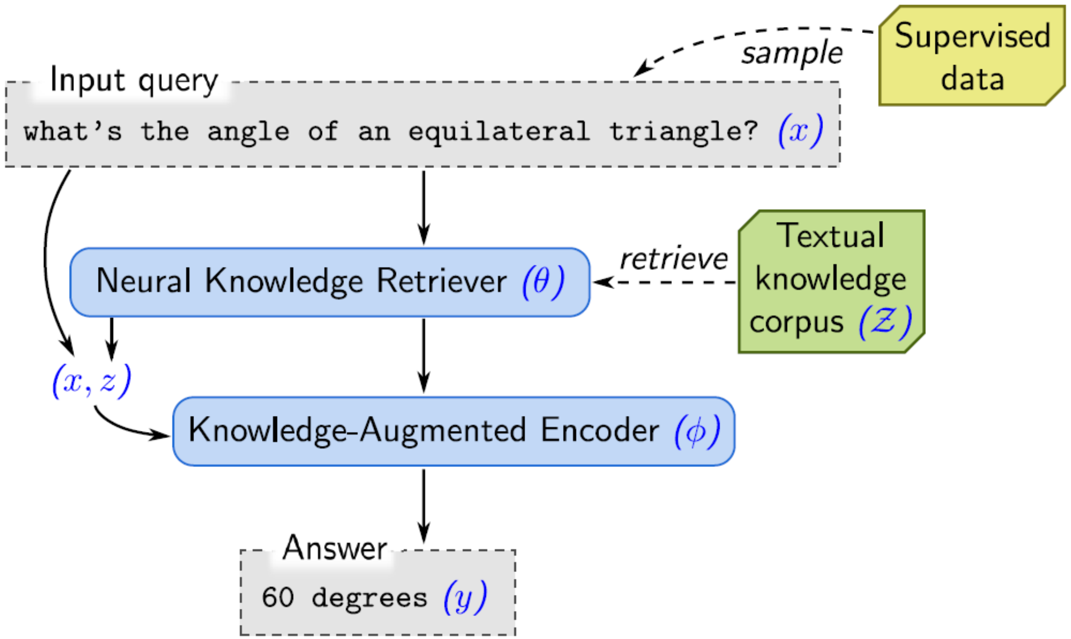
$$p(y \mid z, x) = \prod_{j=1}^{J_x} p(y_j \mid z, x)$$

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

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

RA-LLM Learning: Joint Training

- **REALM**



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

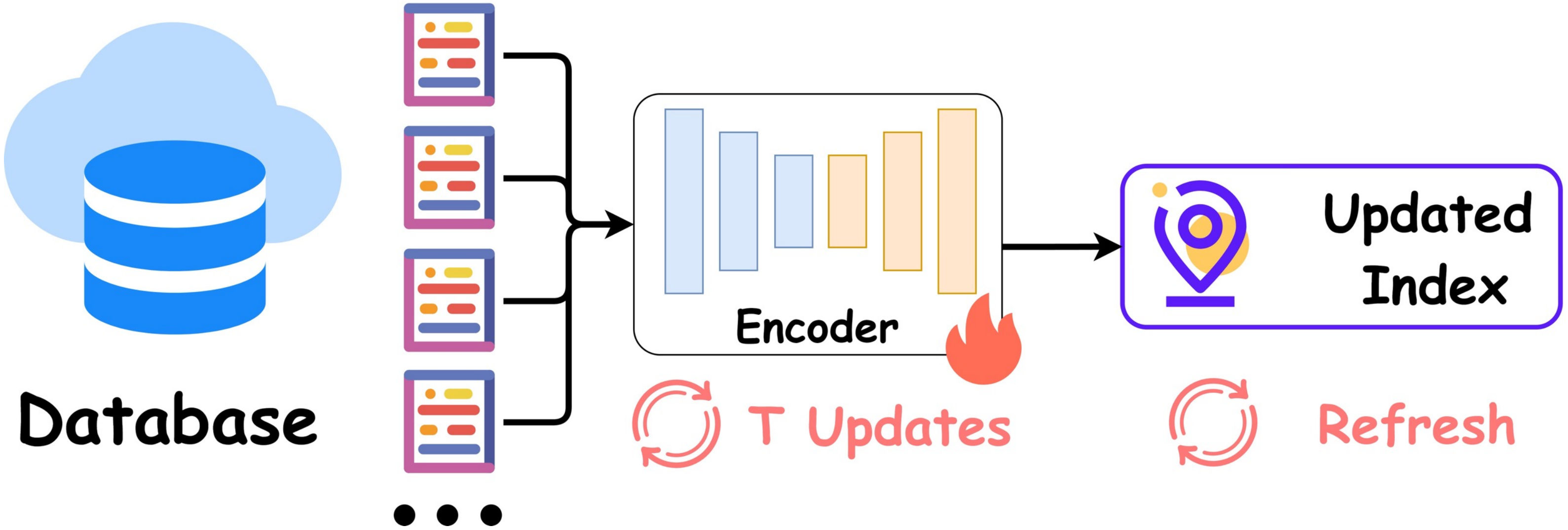
$$h_{\mathtt{START(s)}} = \mathrm{BERT}_{\mathtt{START(s)}}(\mathtt{join}_{\mathrm{BERT}}(x, z_{\mathrm{body}})),$$

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

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

RA-LLM Learning: Joint Training

- **REALM – Asynchronous Index Update**

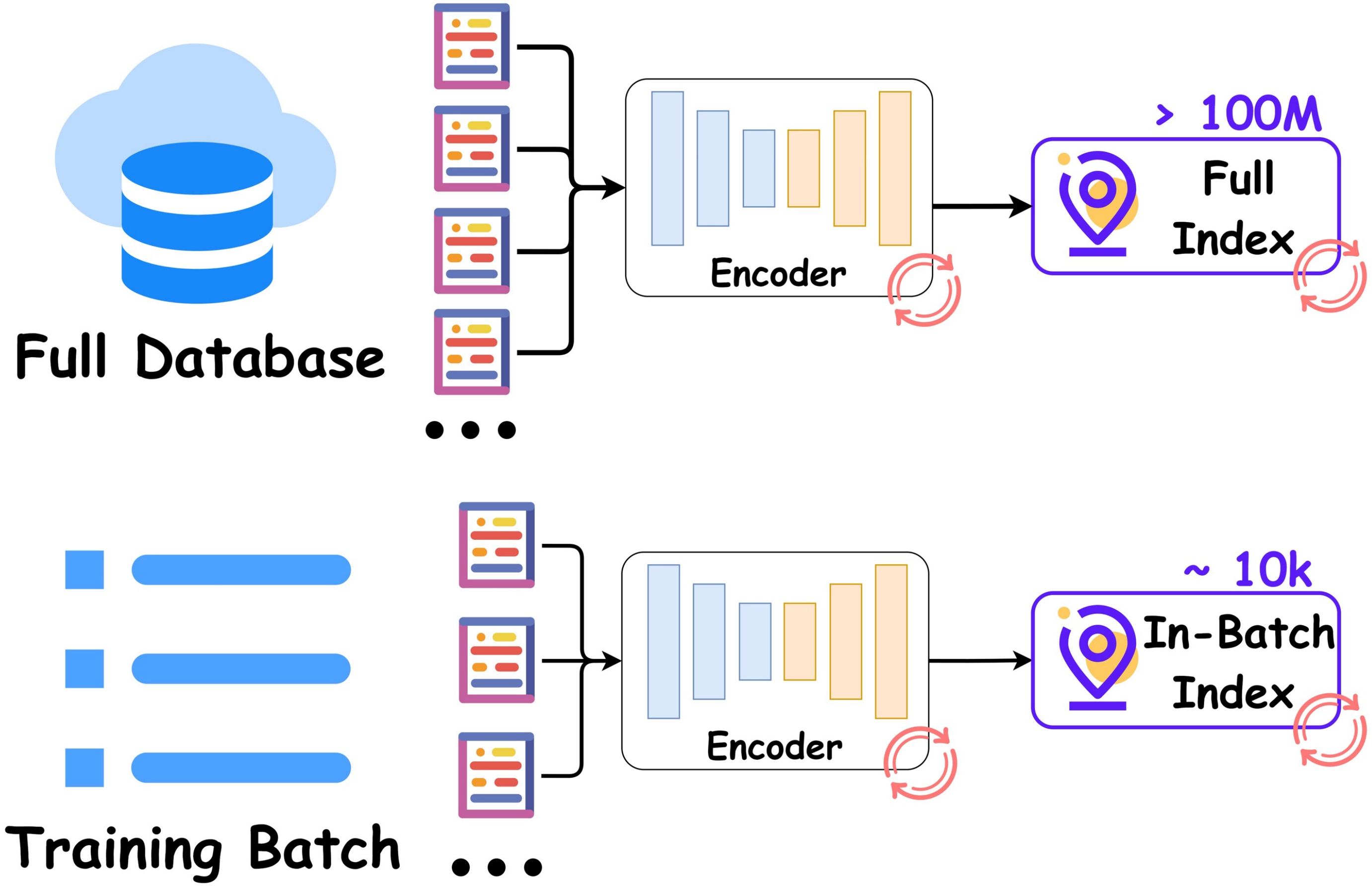


$$f(x, z) = \texttt{Embed}_{\texttt{input}}(x)^{\top}\texttt{Embed}_{\texttt{doc}}(z)$$

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

RA-LLM Learning : Joint Training

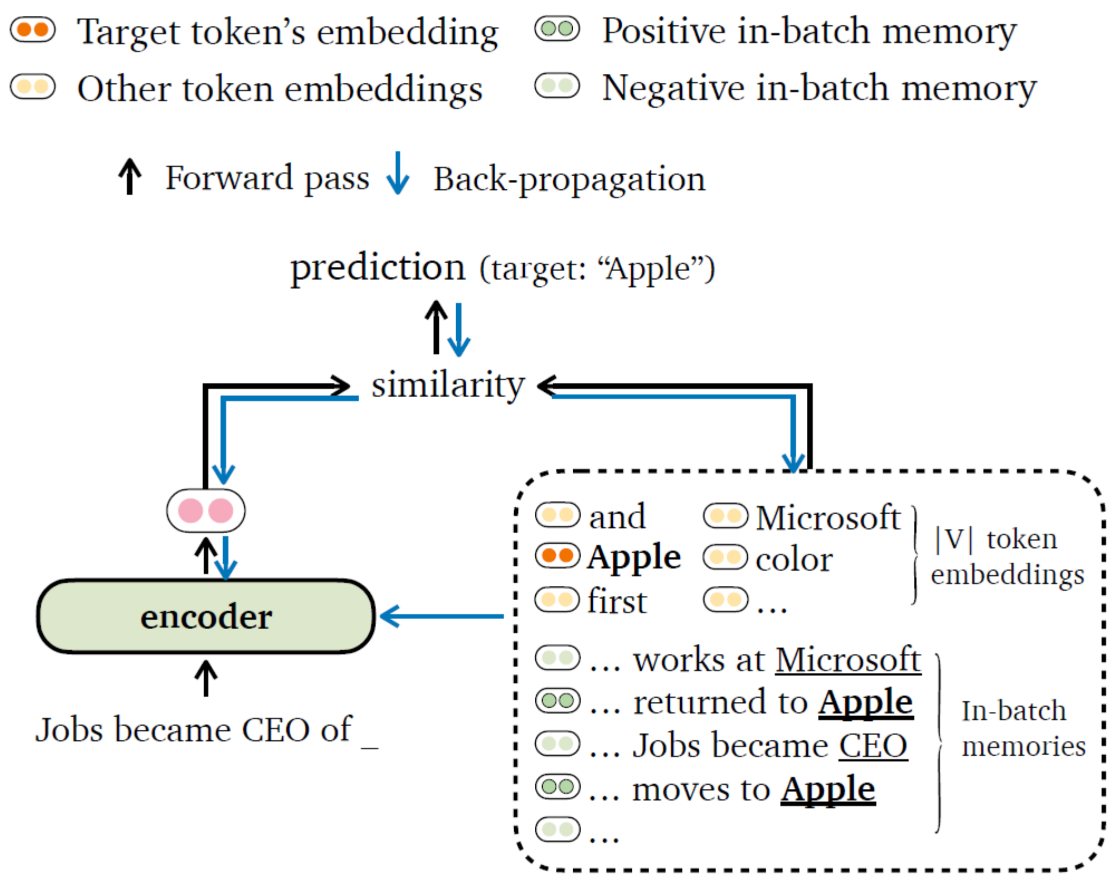
- **TRIME – In-Batch Approximation**



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

RA-LLM 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: $\mathcal{M}_{\text{ext}} = \{(c_j, x_j) \in \mathcal{D}\}.$

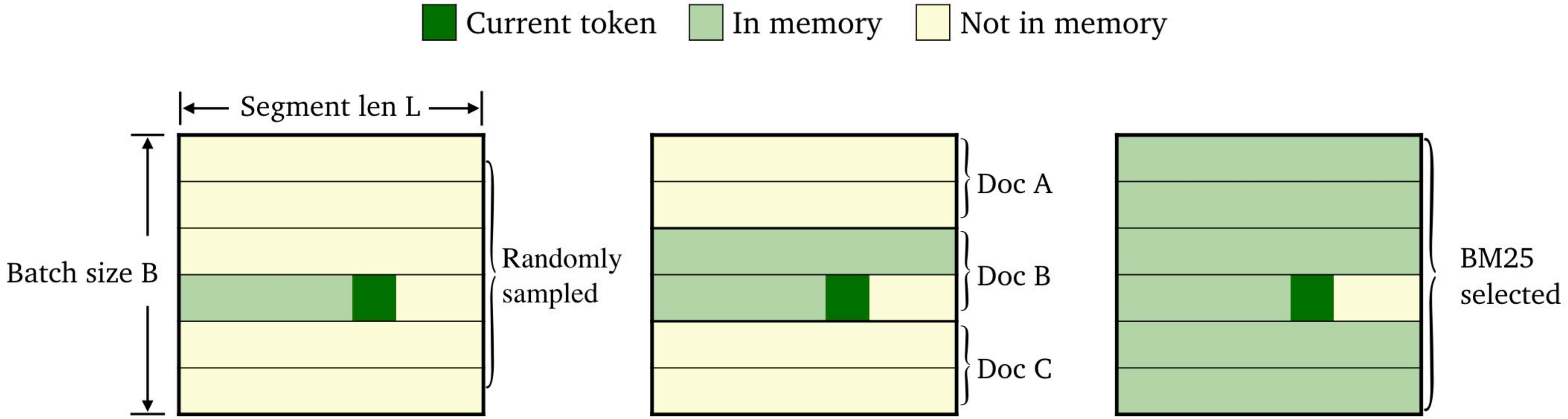
Training Objective:

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

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

RA-LLM Learning : Joint Training

- **TRIME Data Batching Strategy**



(a) Default batching

(b) Batching consecutive segments

(c) Batching lexically similar segments

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 Language Models (RA-LLMs) (Dr. Wenqi Fan)
- **Part 2: Architecture** of RA-LLMs and **Main Modules** (Dr. Yujuan Ding)
- **Part 3: Learning** Approach of RA-LLMs (Liangbo Ning)
- **Part 4: Applications of RA-LLMs (Shijie Wang)**
- **Part 5: Challenges and Future Directions** of RA-LLMs (Dr. Wenqi Fan)

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

Website

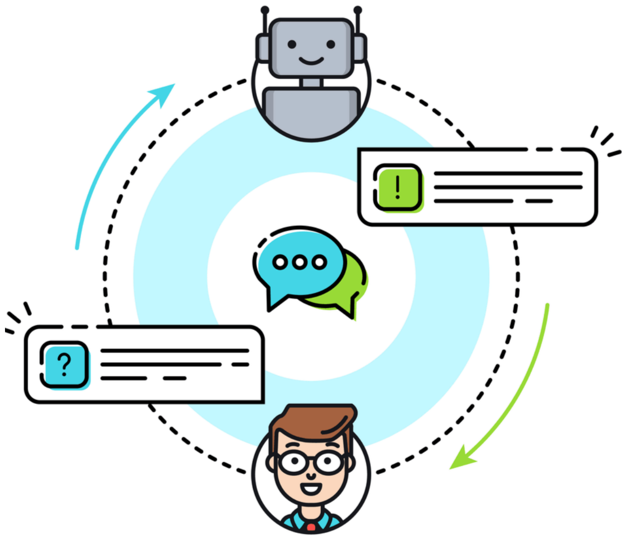
PART 4: Application of RA-LLMs

Presenter
Shijie Wang
HK PolyU

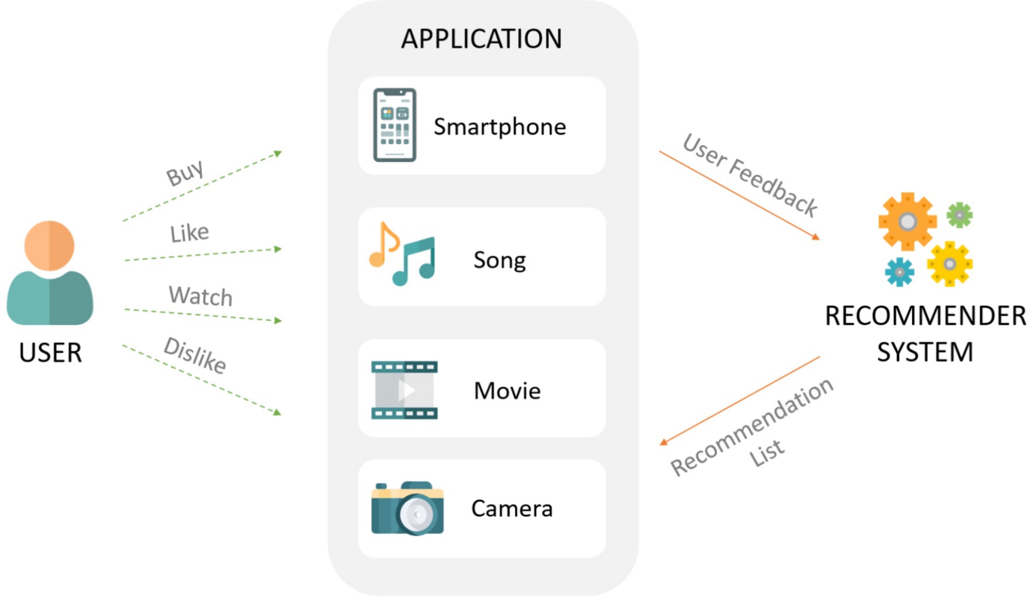
- **NLP applications**
- **Downstream tasks**
- **Domain-specific applications**

RA-LLM Applications

- **Various applications**



Chatbots



Recommendation

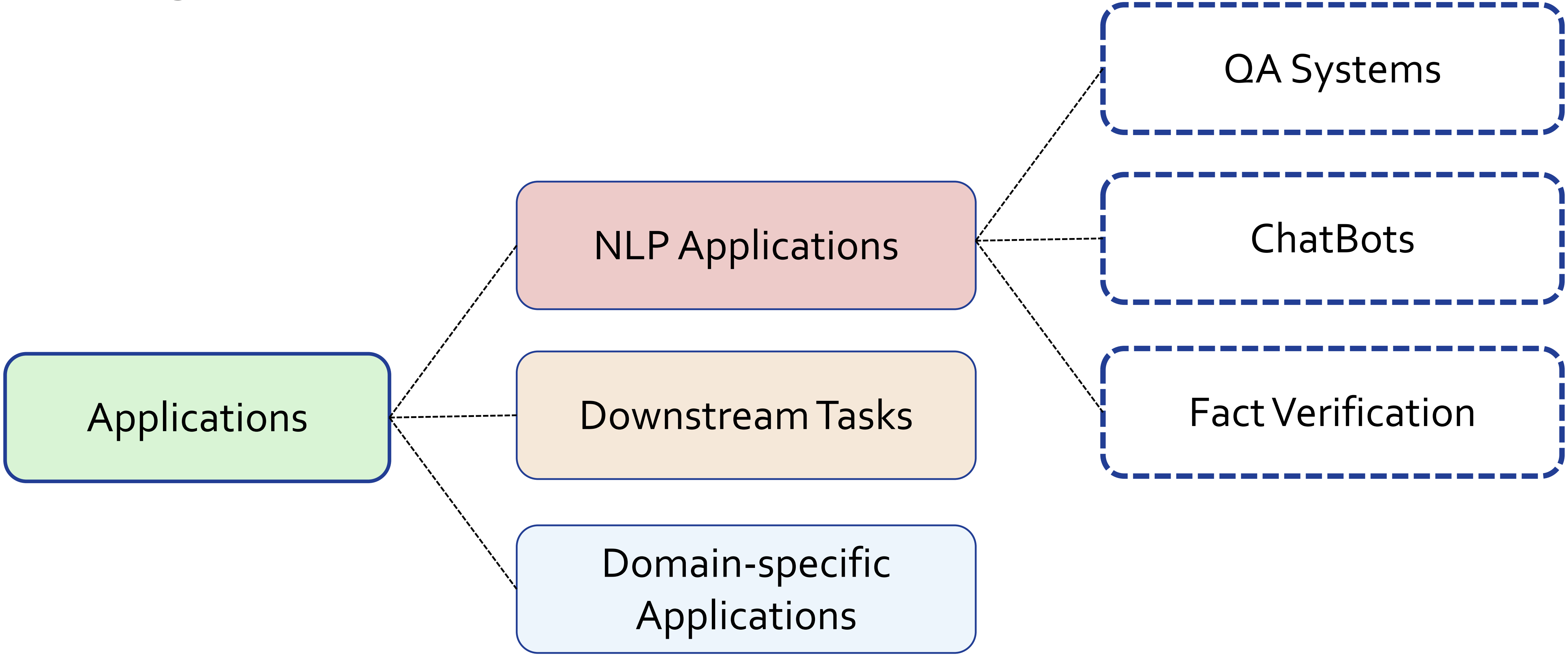
AI for Science

…

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

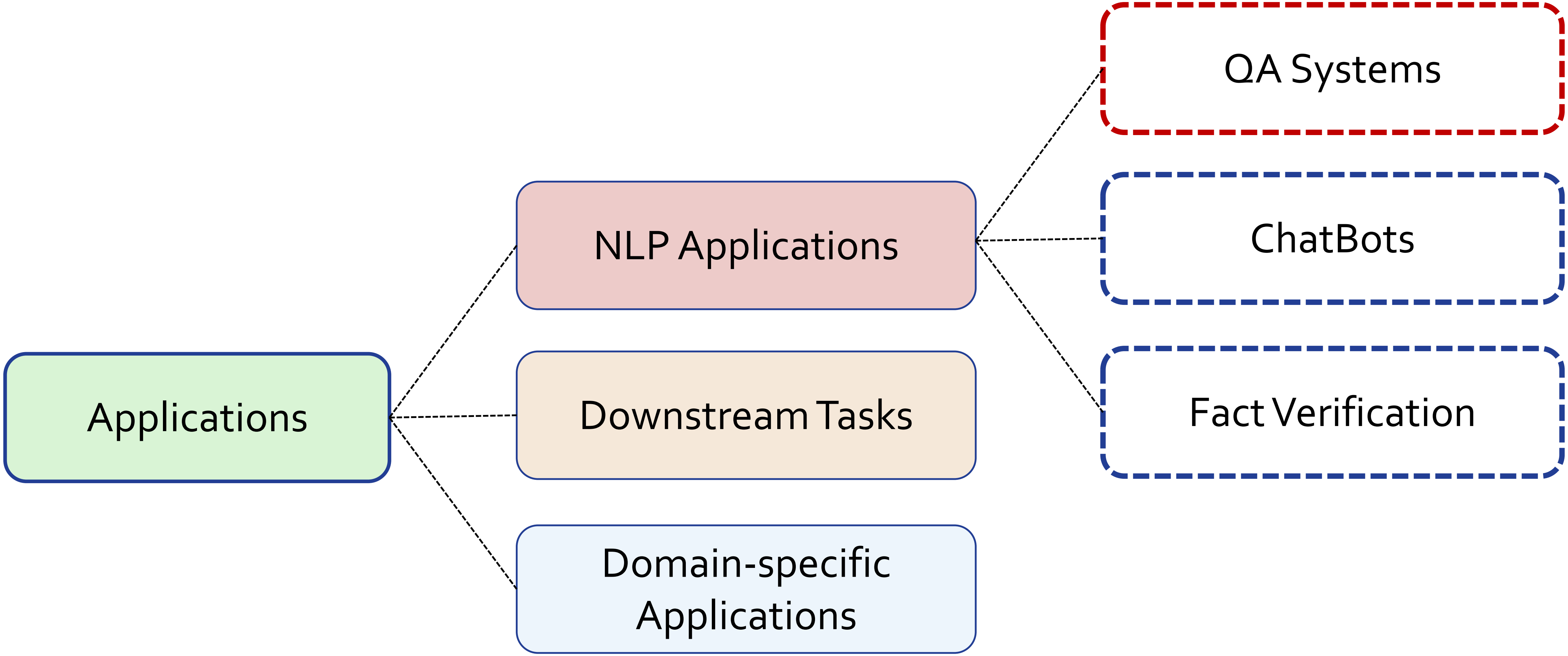
RA-LLM Applications: NLP Applications

- **Categories**



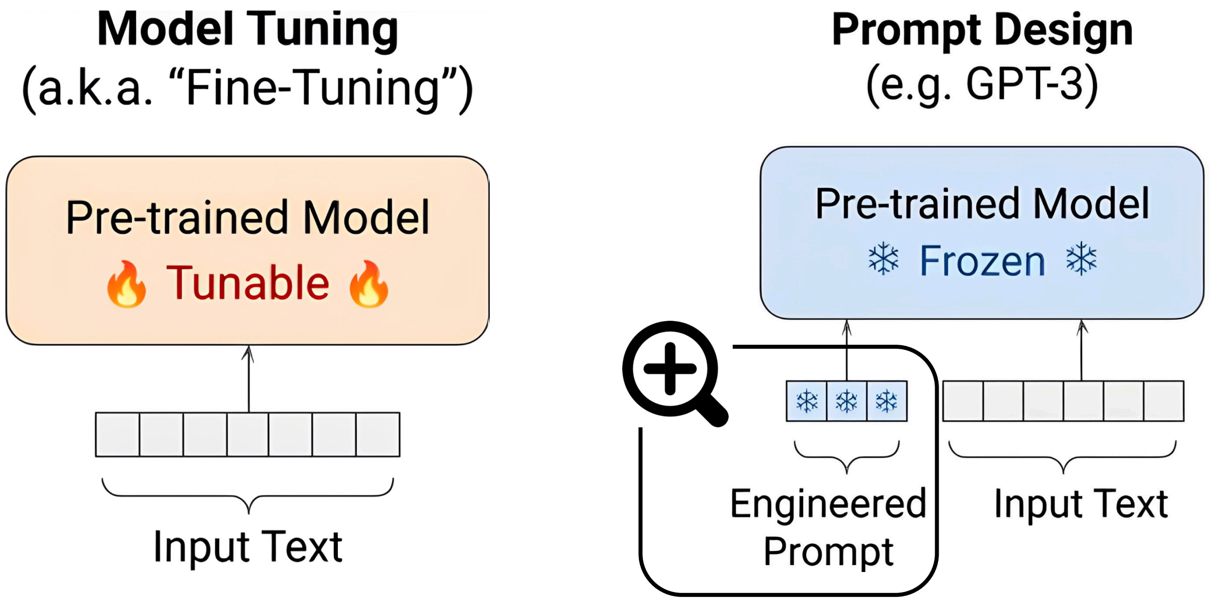
RA-LLM Applications: NLP Applications

- **QA Systems**



RA-LLM Applications: QA Systems

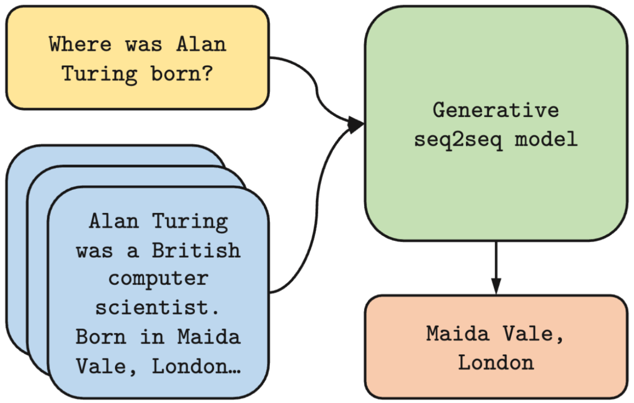
- **QA systems**
  - Challenges:
    - Open-domain QA
    - Domain-specific QA
  - How to solve?
    - Fine-tuning
    - Prompting

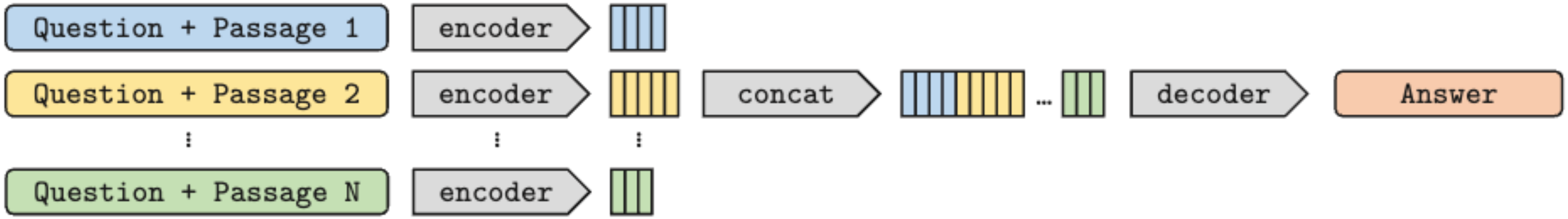


RA-LLM Applications: QA Systems

- **Retrieves for open-domain QA**

  Retrieves support text passages from an external source of knowledge

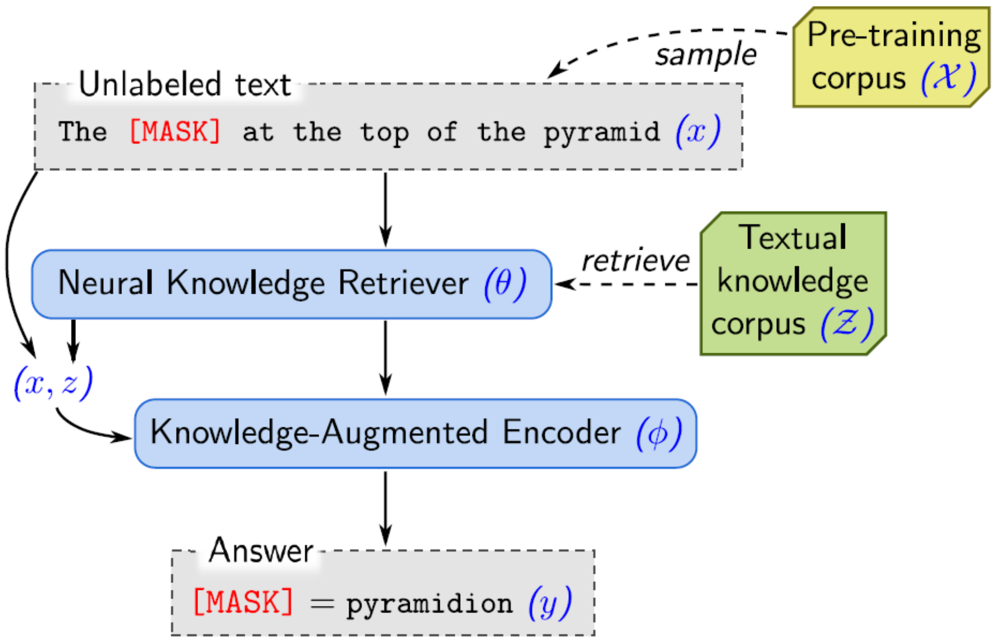


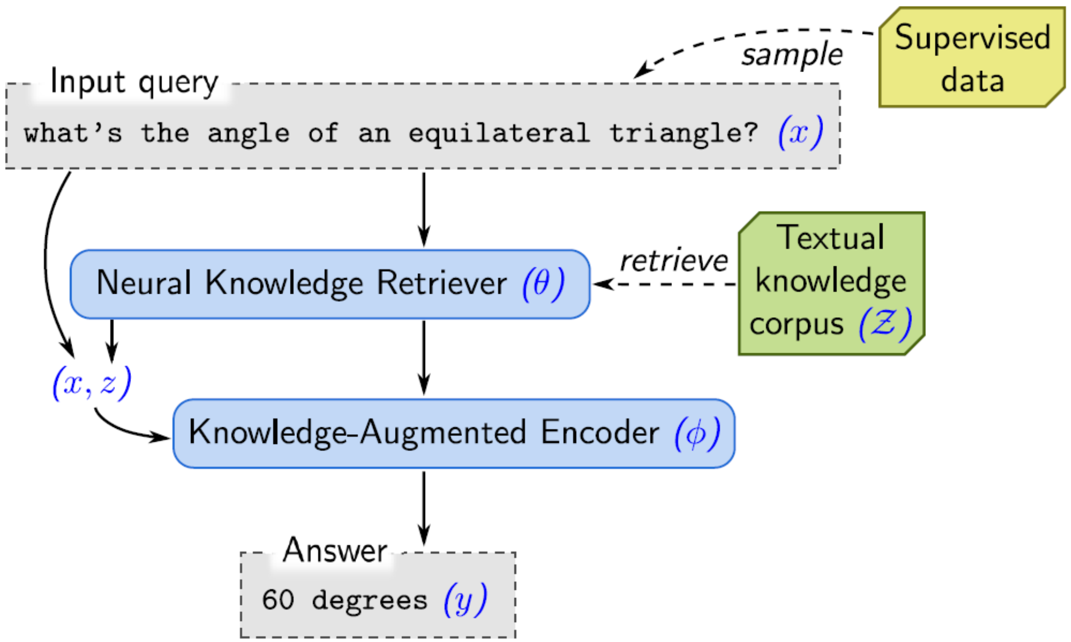


Izacard et al., 2021. "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering"

RA-LLM Applications: QA Systems

- **REALM**

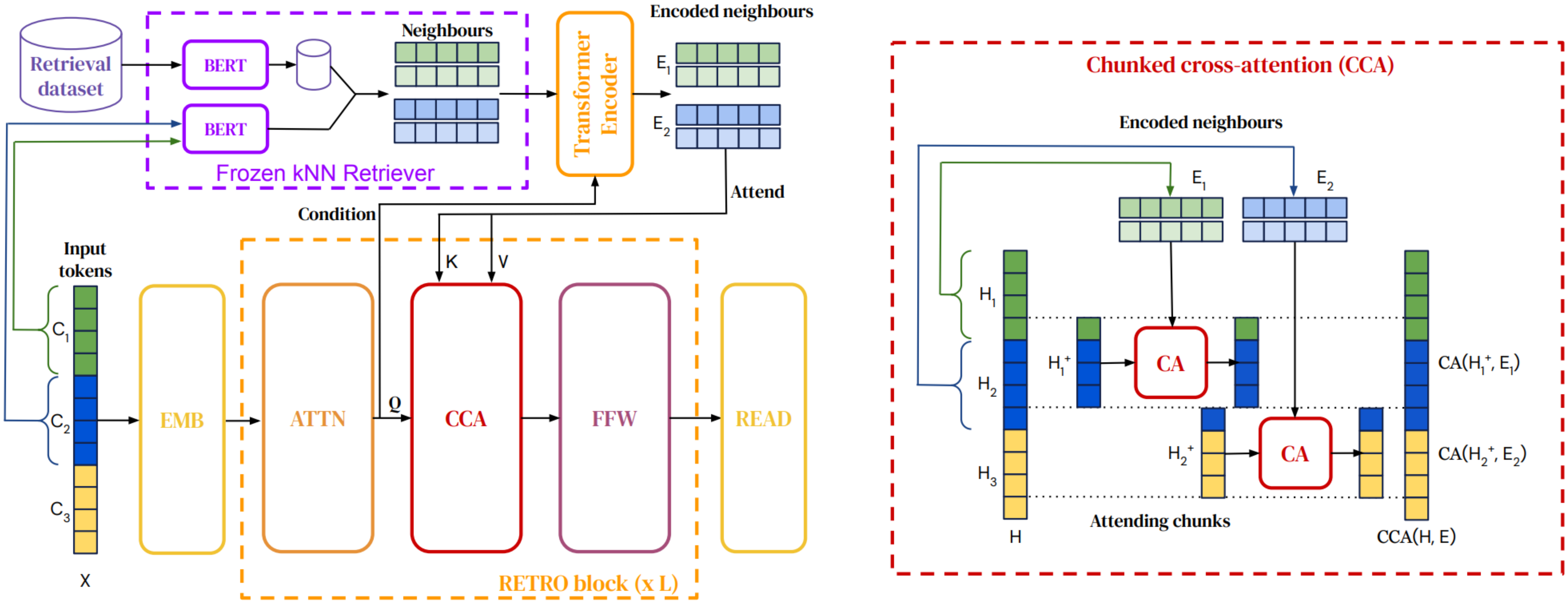




Guu et al., 2020. "REALM: Retrieval-Augmented Language Model Pre-Training"

RA-LLM Applications: QA Systems

- **RETRO (Retrieval-enhanced transformer)**

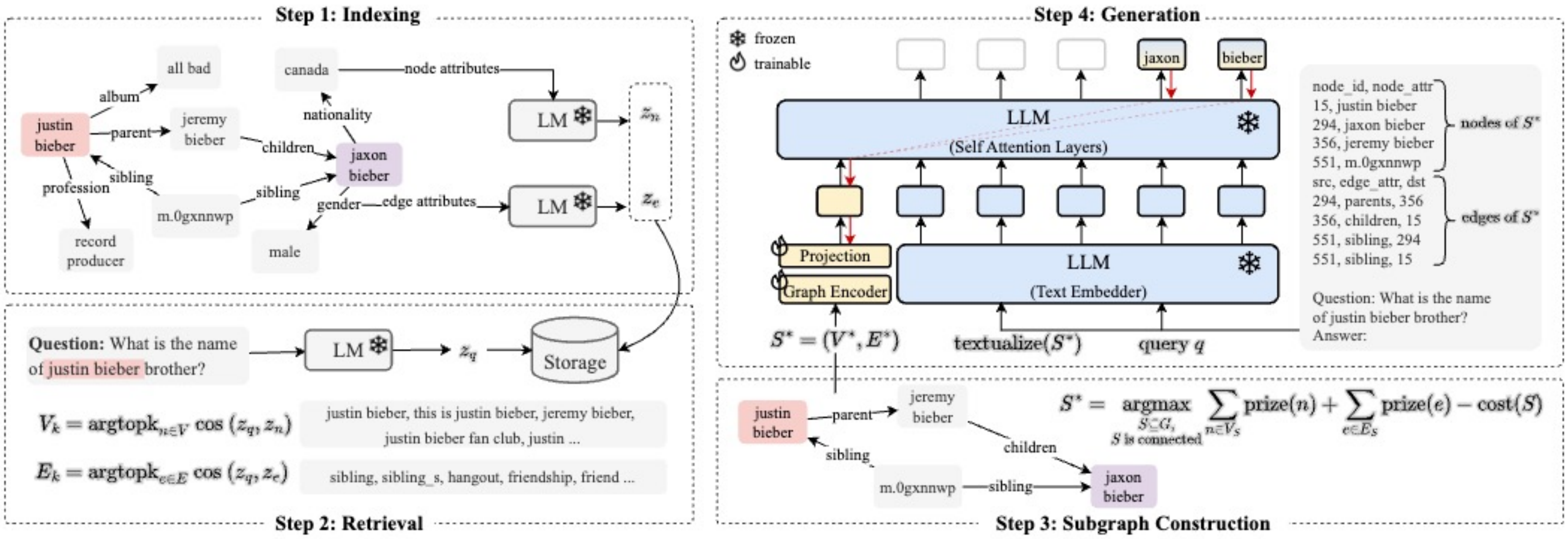


Borgeaud et al., 2022. “Improving Language Models By Retrieving From Trillions Of Tokens”

RA-LLM Applications: QA Systems

- **G-Retriever**

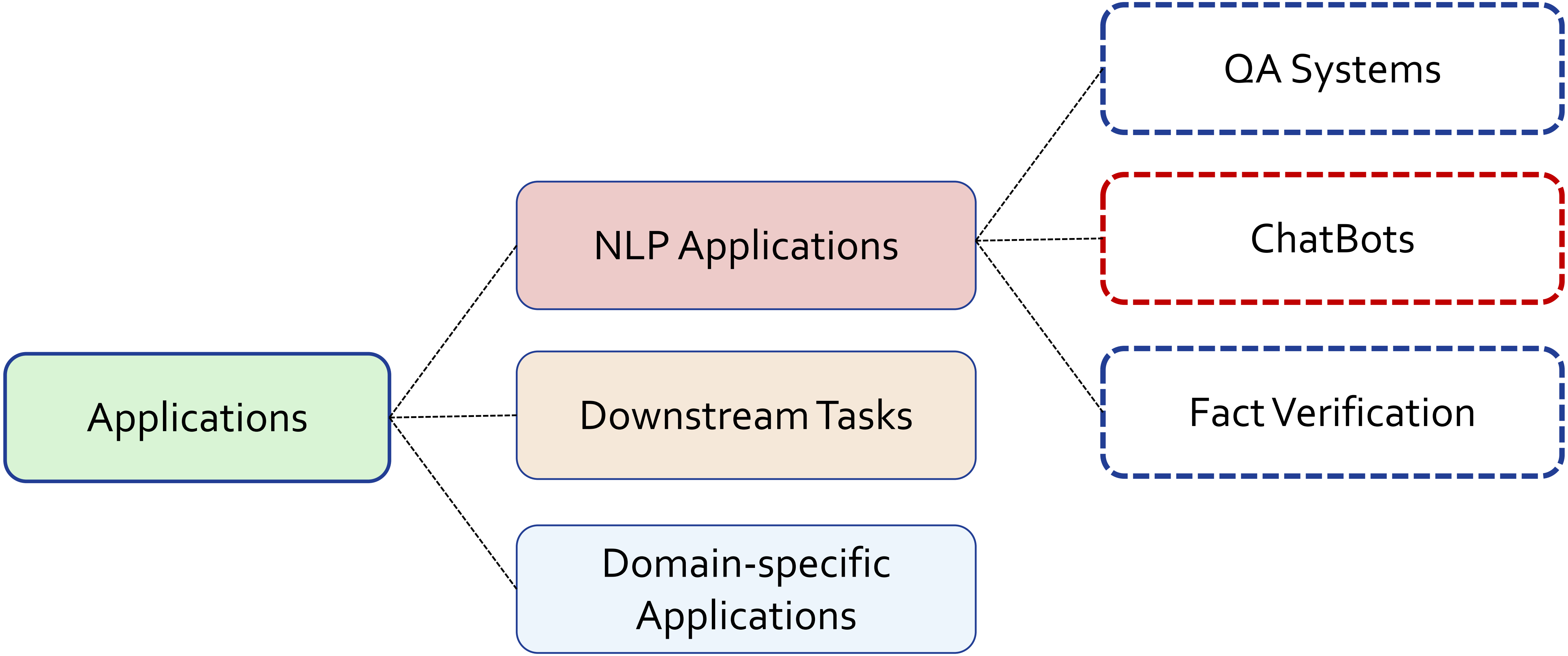
  Retrieves from knowledge graph for question-answering



He, et al. "G-retriever: Retrieval-augmented generation for textual graph understanding and question answering."

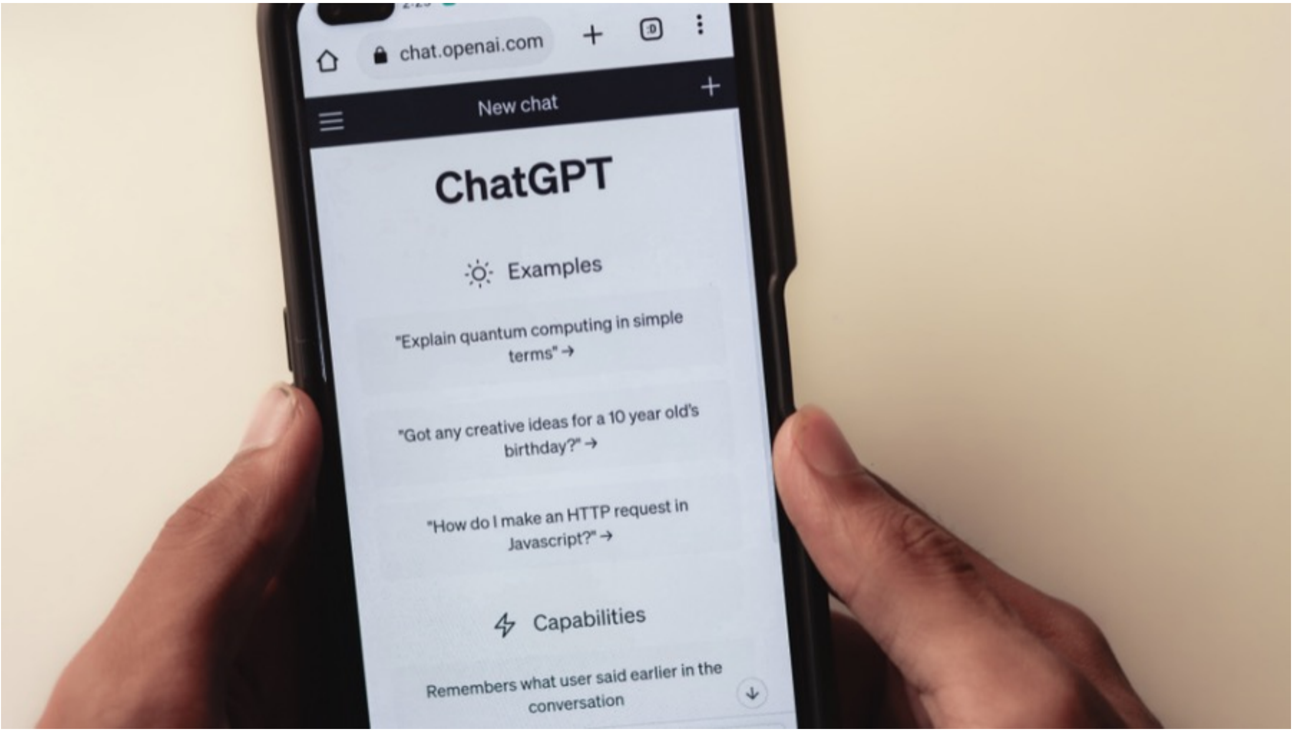
RA-LLM Applications: NLP Applications

- **ChatBots**



RA-LLM Applications: Chatbots

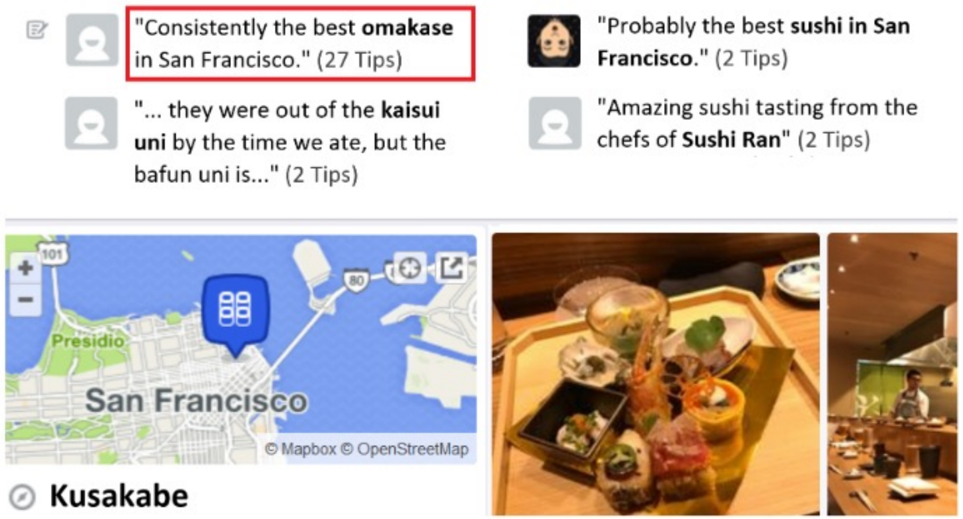
- **ChatBots**



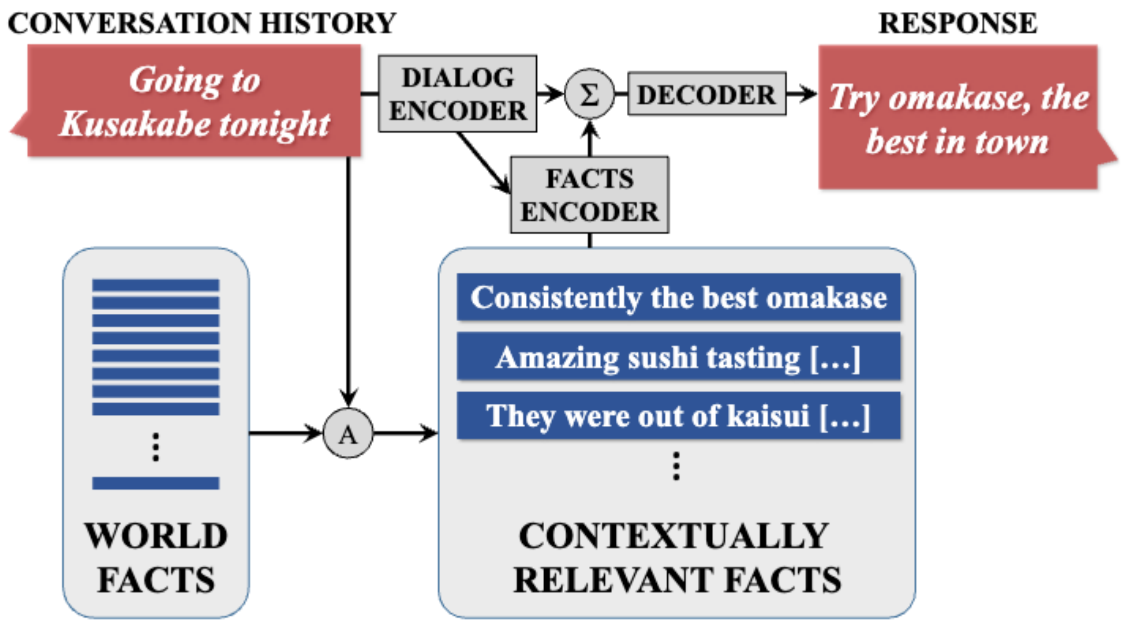
https://www.forbes.com/advisor/in/business/software/what-is-a-chatbot/

RA-LLM Applications: Chatbots

- **Knowledge-grounded model**



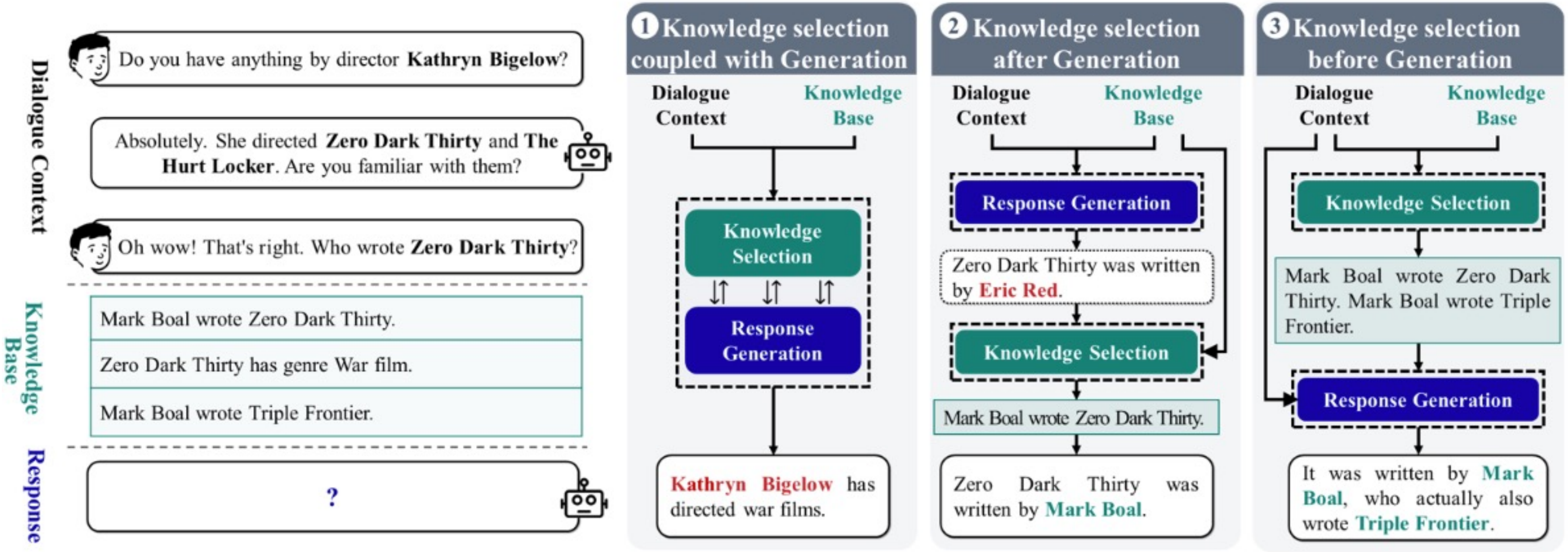
User input: *Going to Kusakabe tonight.*
Neural model: Have a great time!
Human: You'll love it! Try omasake, the best in town.



Ghazvininejad et al., 2018. "A Knowledge-Grounded Neural Conversation Model"

RA-LLM Applications: Chatbots

- **GATE**

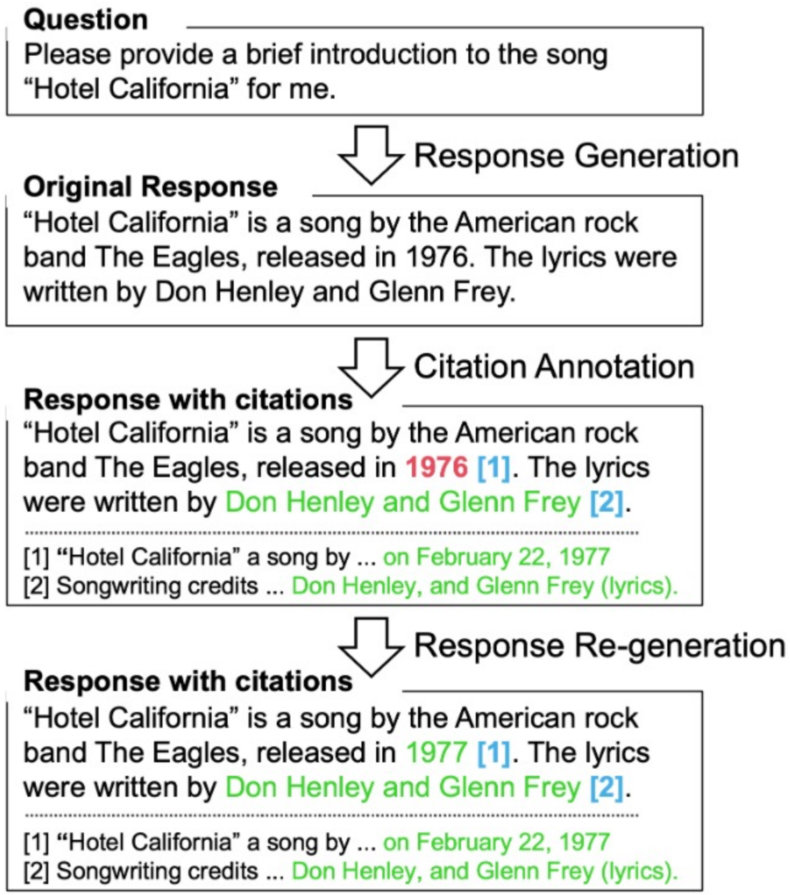


(a) Example of knowledge-grounded dialogue (b) Three categories of knowledge selection

Qin et al., 2023. "Well Begun is Half Done: Generator-agnostic Knowledge Pre-Selection for Knowledge-Grounded Dialogue"

RA-LLM Applications: Chatbots

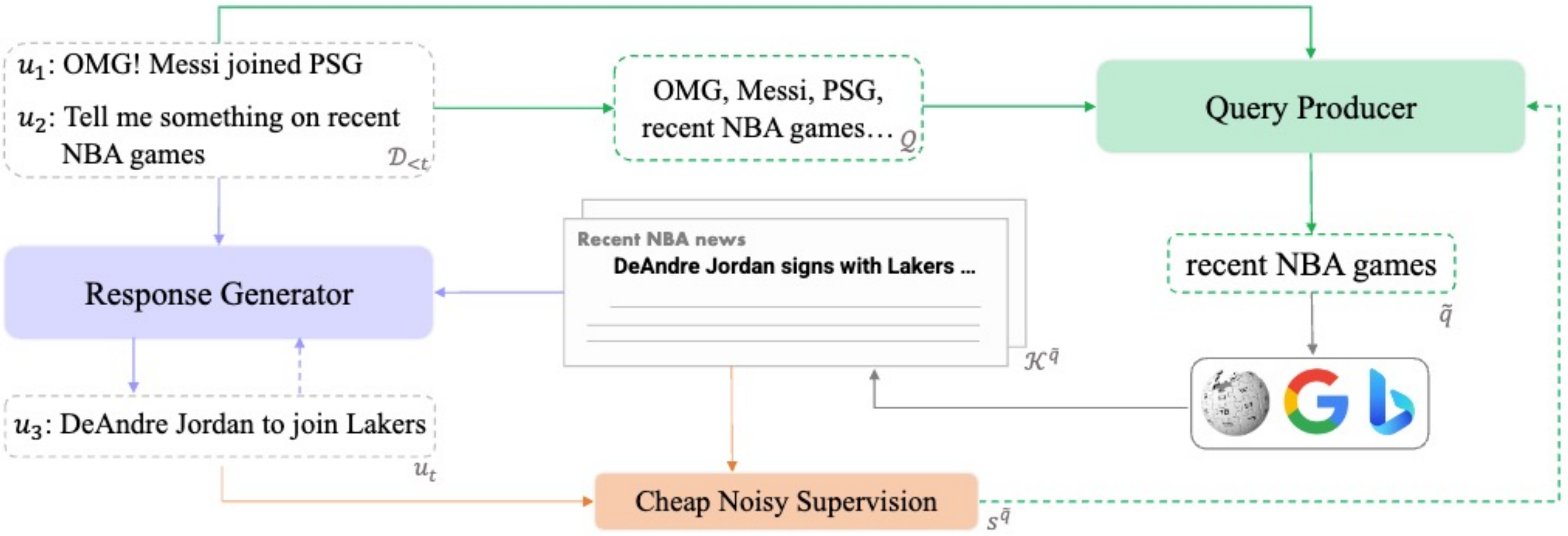
- **CEG**



Li et al., 2024. "Citation-Enhanced Generation for LLM-based Chatbots"

RA-LLM Applications: Chatbots

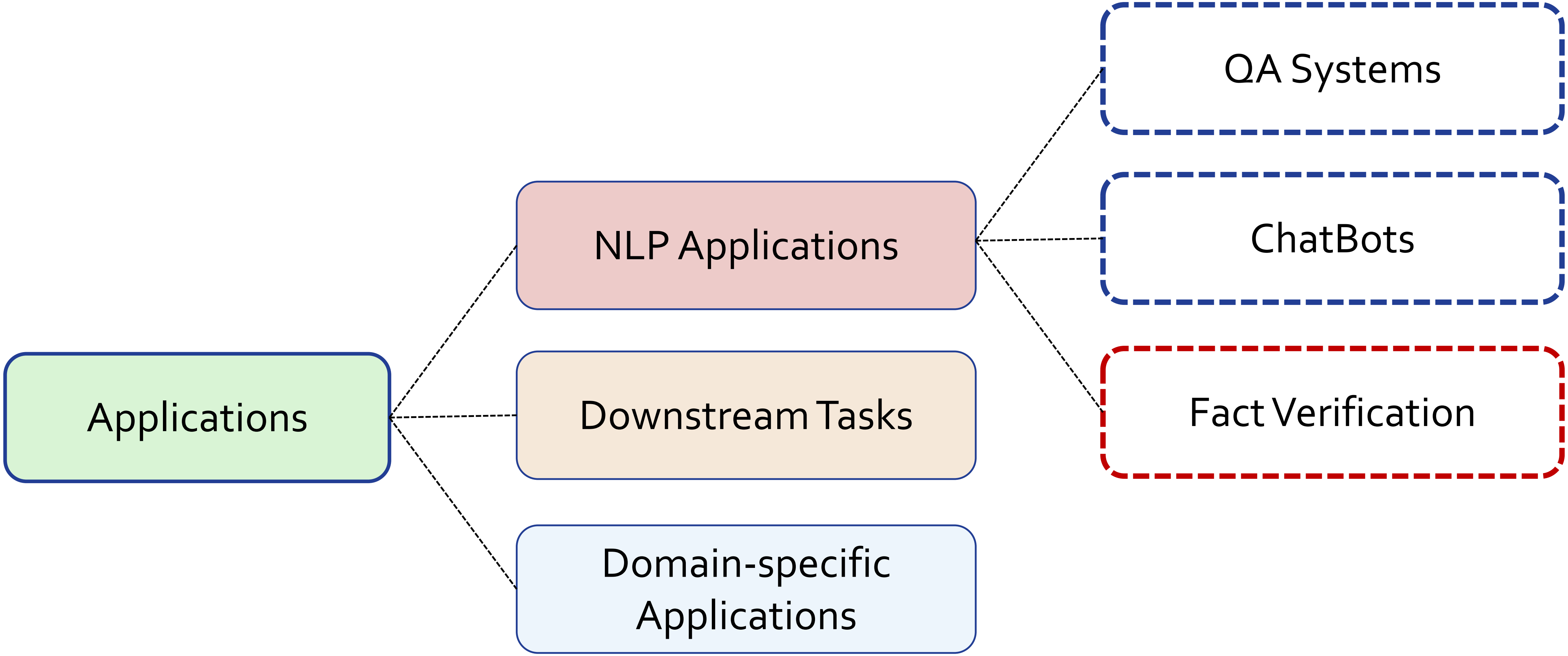
- **Search-engine-augmented chatbots**



Wang et al., 2023. “Search-Engine-augmented Dialogue Response Generation with Cheaply Supervised Query Production”

RA-LLM Applications: NLP Applications

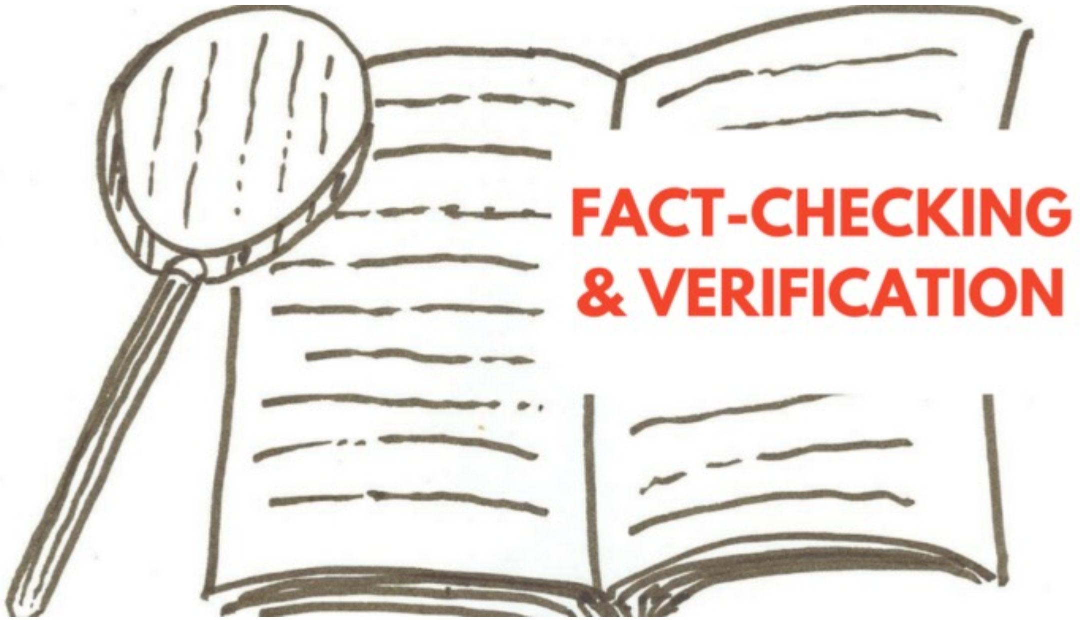
- **Fact verification**



RA-LLM Applications: Fact Verification

- **Fact verification**

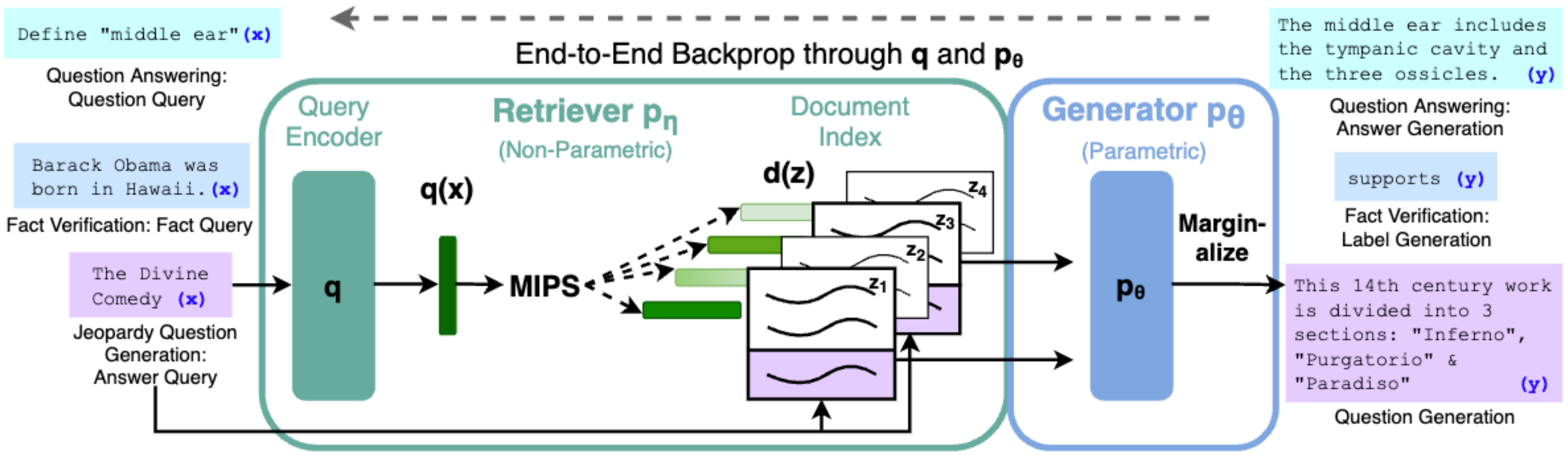
  Fact Verification is a critical task in verifying the accuracy and reliability of information



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

RA-LLM Applications: Fact Verification

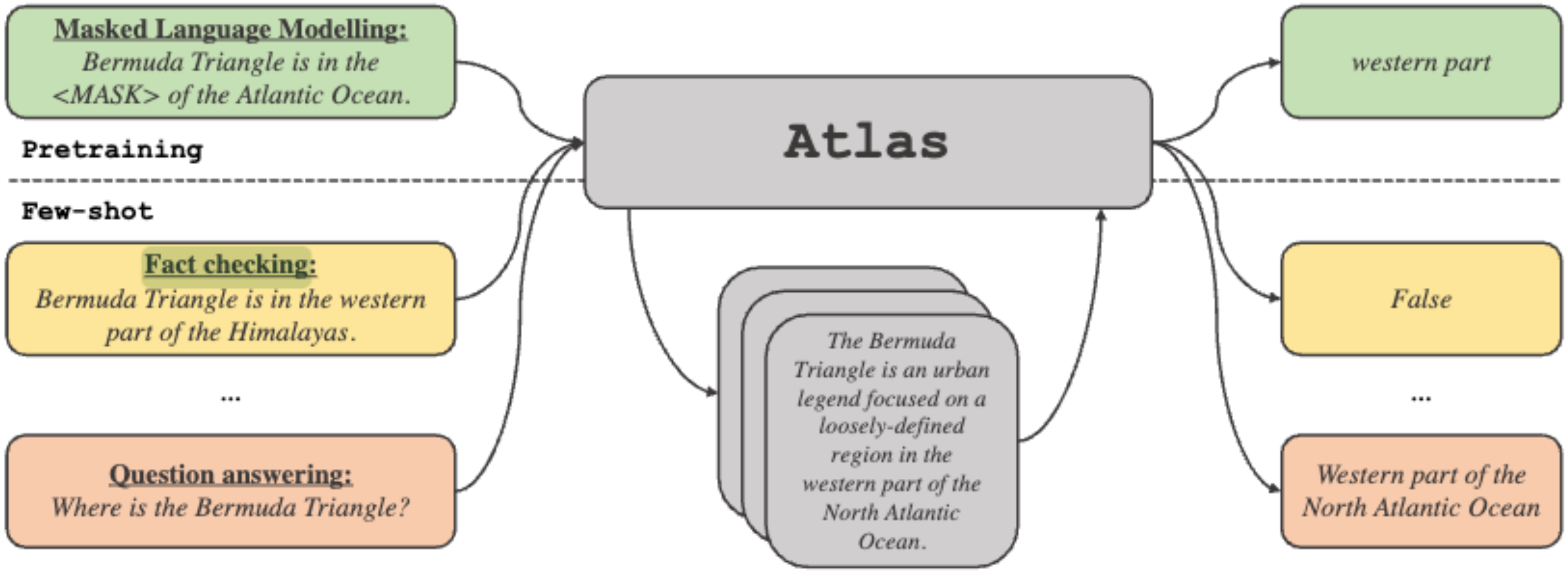
- **Fact verification**



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

RA-LLM Applications: Fact Verification

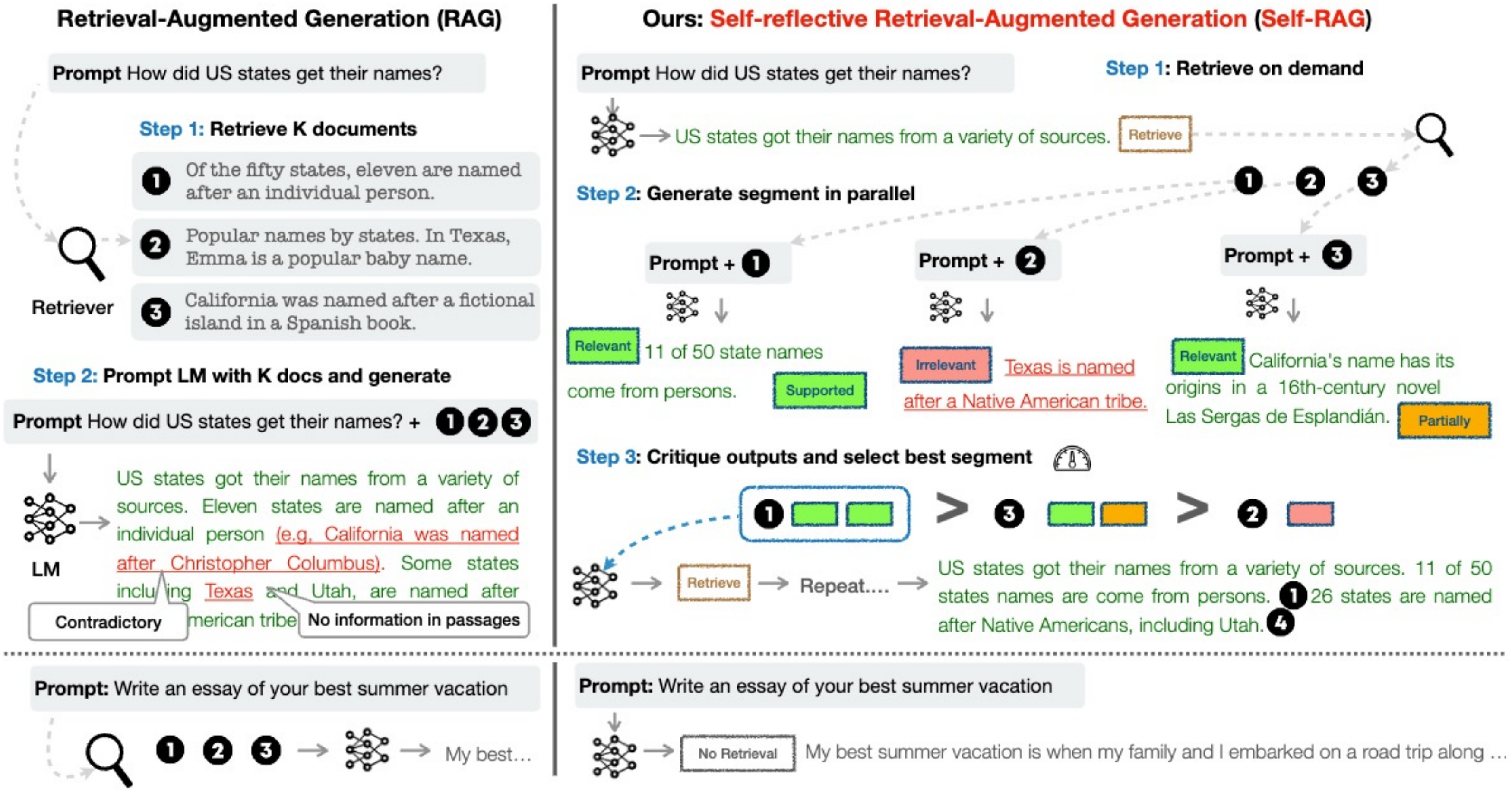
- **Fact verification**
  - Fact verification is usually together with other NLP tasks (such as Q & A)
  - ATLAS:



Izacard et al., 2023. "Atlas: Few-shot Learning with Retrieval Augmented Language Models"

RA-LLM Applications: Fact Verification

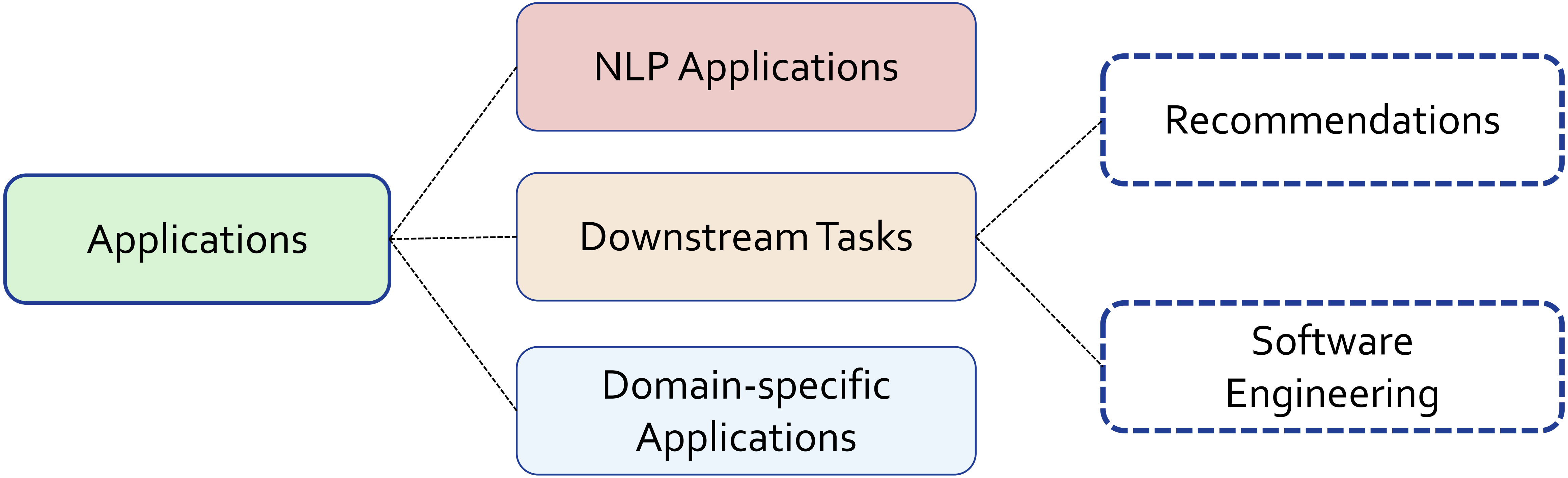
- **Self-RAG**



Asai et al., 2023. “Self-rag: Learning To Retrieve, Generate, And Critique Through Self-reflection”

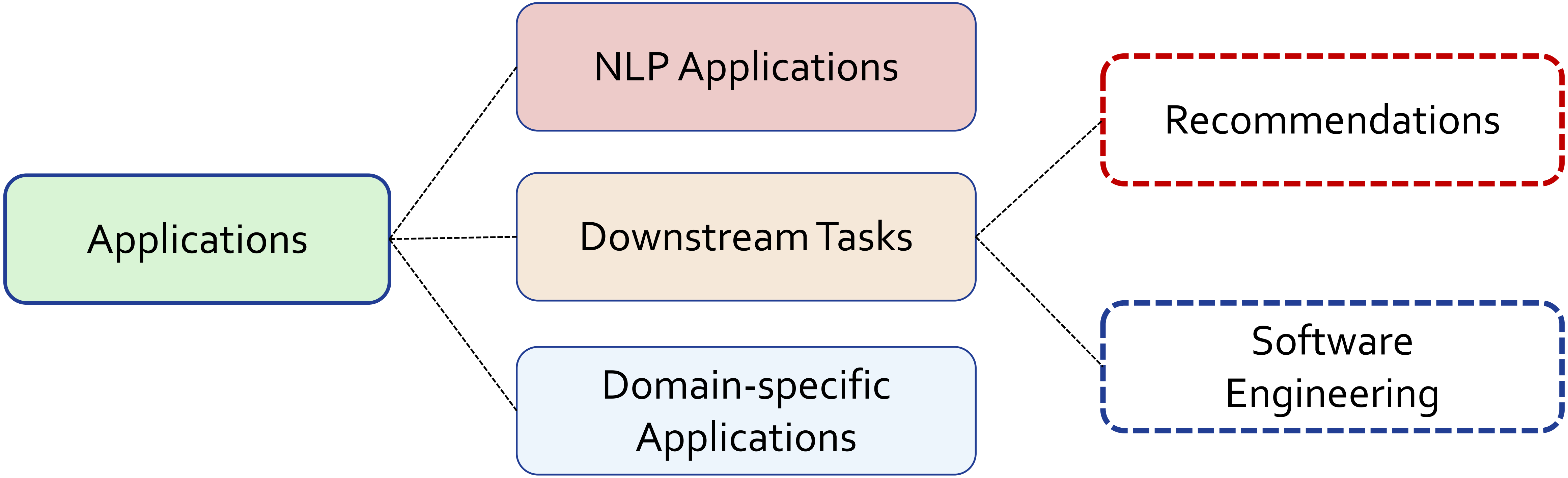
RA-LLM Applications: Downstream Tasks

- **Downstream tasks**



RA-LLM Applications: Recommendations

- **Recommendations**



RA-LLM Applications: Recommendations

- **Recommendations**

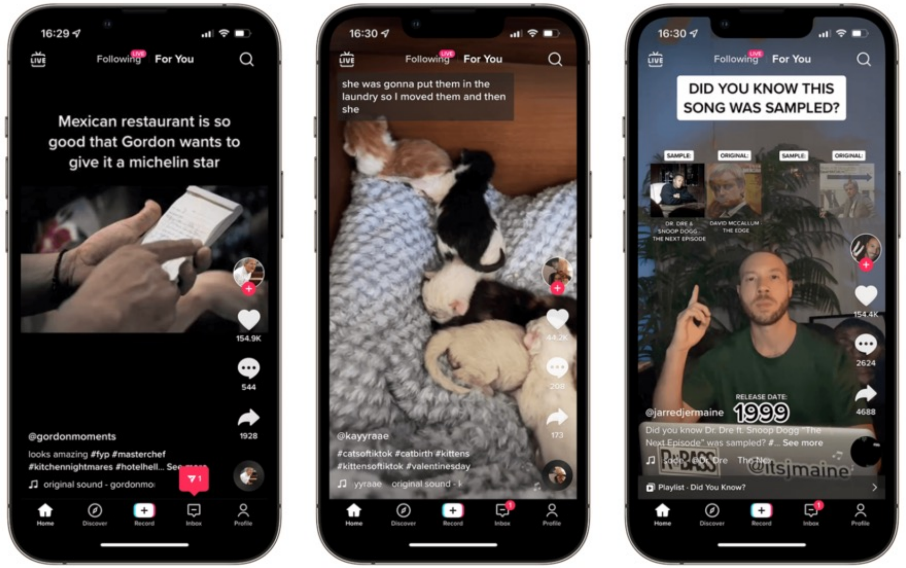
  Recommendation has been widely applied in online services

YouTube

TikTok

Google News

News/Video/Image Recommendation



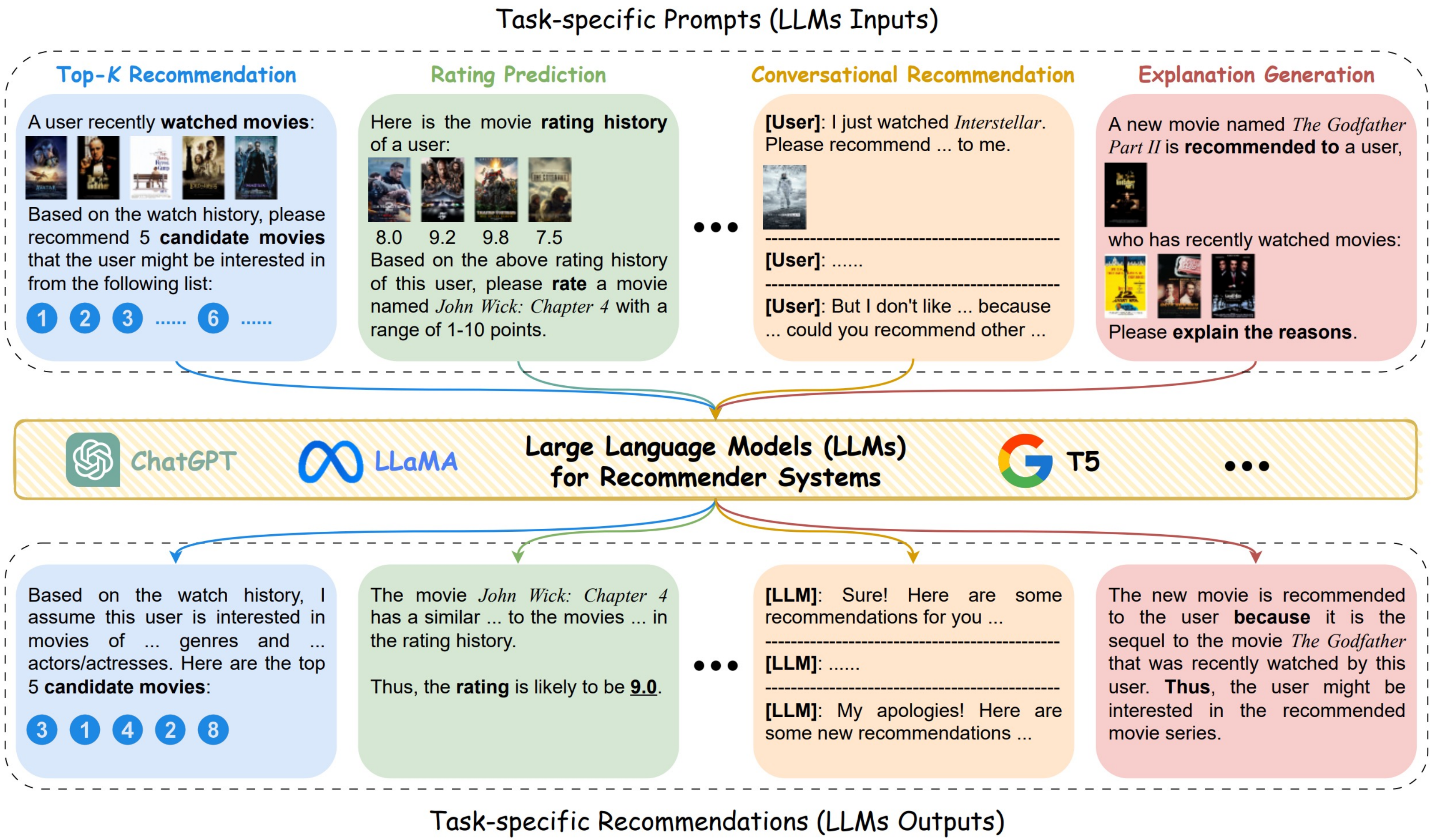
<u>TikTok's recommendation algorithm</u>

Top 10 Global Breakthrough Technologies in 2021

MIT Technology Review

RA-LLM Applications: Recommendations

- **LLMs in recommendations**



RA-LLM Applications: Recommendations

- **Conventional item-based LLM reasoning process**

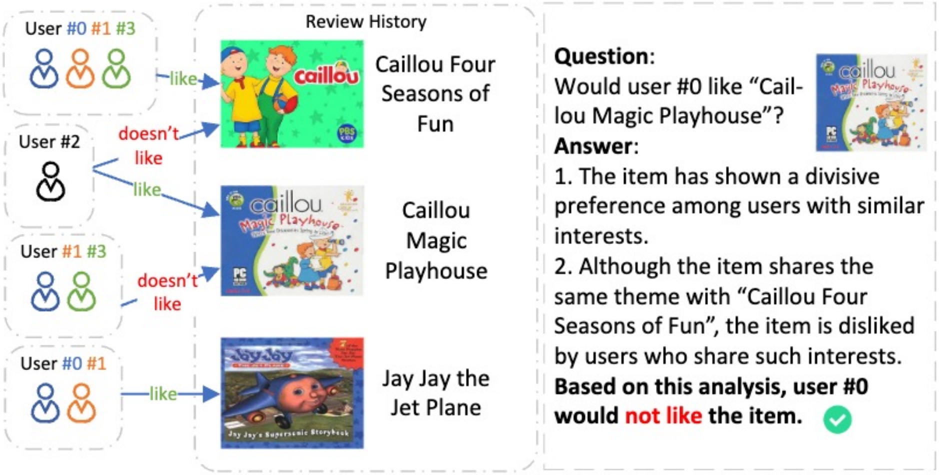
(a) Conventional item-based [16, 42] LLM reasoning process.

Wu et al., 2024. "CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation"

RA-LLM Applications: Recommendations

- **Collaborative retrieval augmented LLM reasoning process**

(a) Conventional item-based [16, 42] LLM reasoning process.

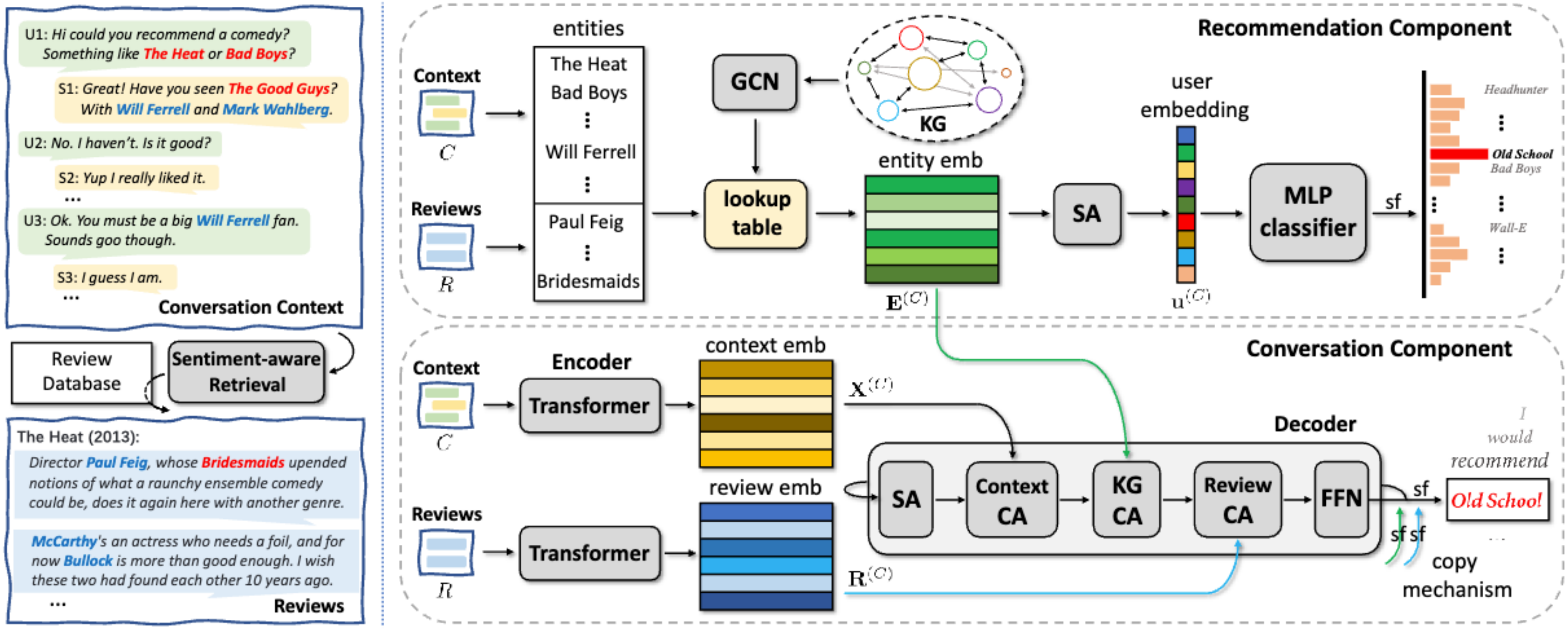


(b) Collaborative Retrieval Augmented LLM reasoning process.

Wu et al., 2024. "CoRAL: Collaborative Retrieval-Augmented Large Language Models Improve Long-tail Recommendation"

RA-LLM Applications: Recommendations

- **Retrieval from the reviews**

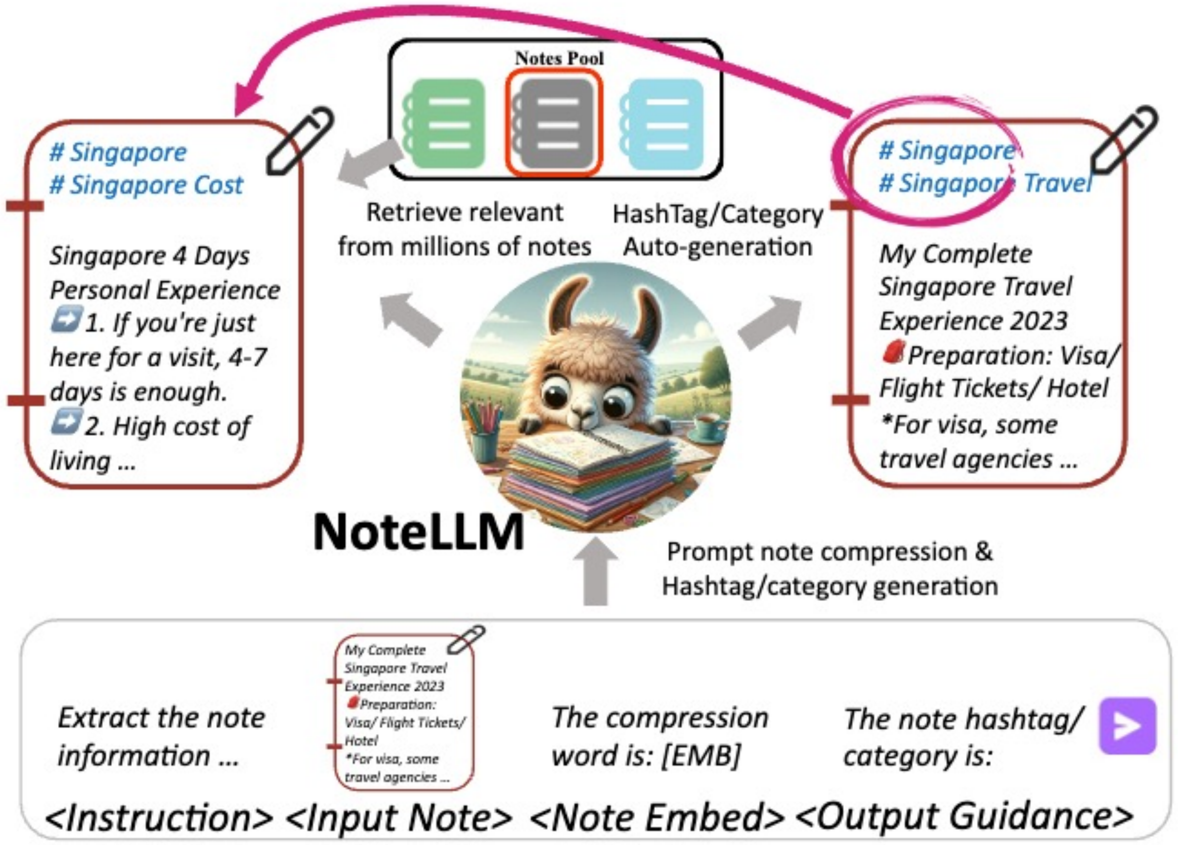


Lu et al., 2021. “RevCore: Review-augmented Conversational Recommendation”

RA-LLM Applications: Recommendations

- **Retrieval from the notes**

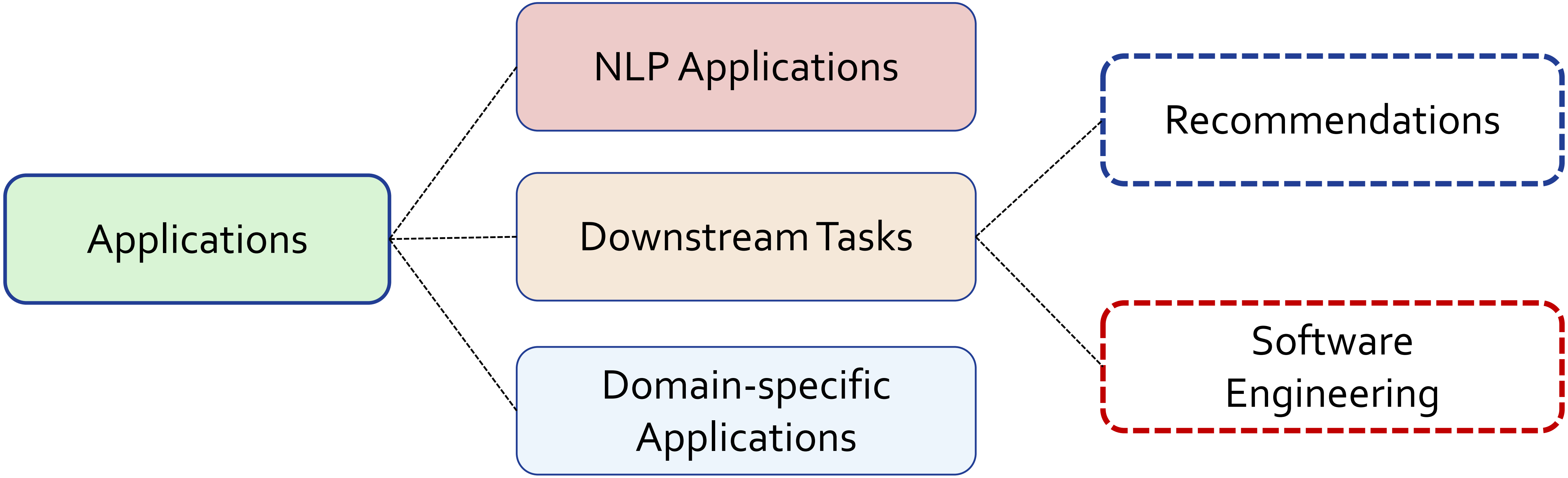
NoteLLM:



Zhang et al., 2024. "NoteLLM: A Retrievable Large Language Model for NoteRecommendation"

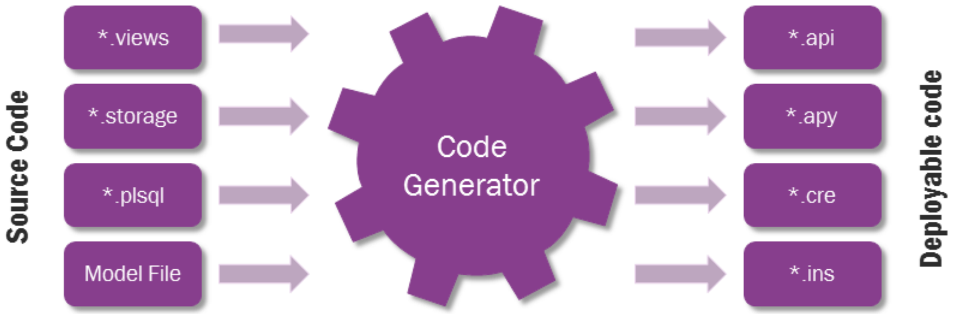
RA-LLM Applications: Software Engineering

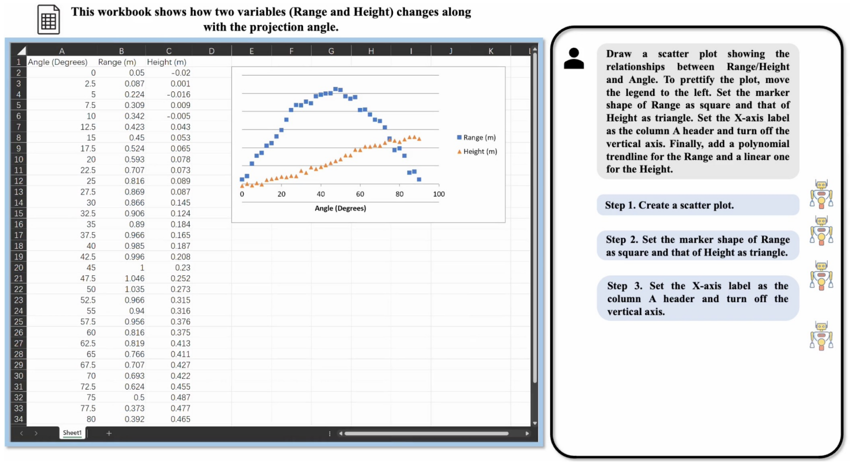
- **Software engineering**



RA-LLM Applications: Software Engineering

- **Software engineering:**
  - Code generation
  - Program repair
  - Table processing
  - …

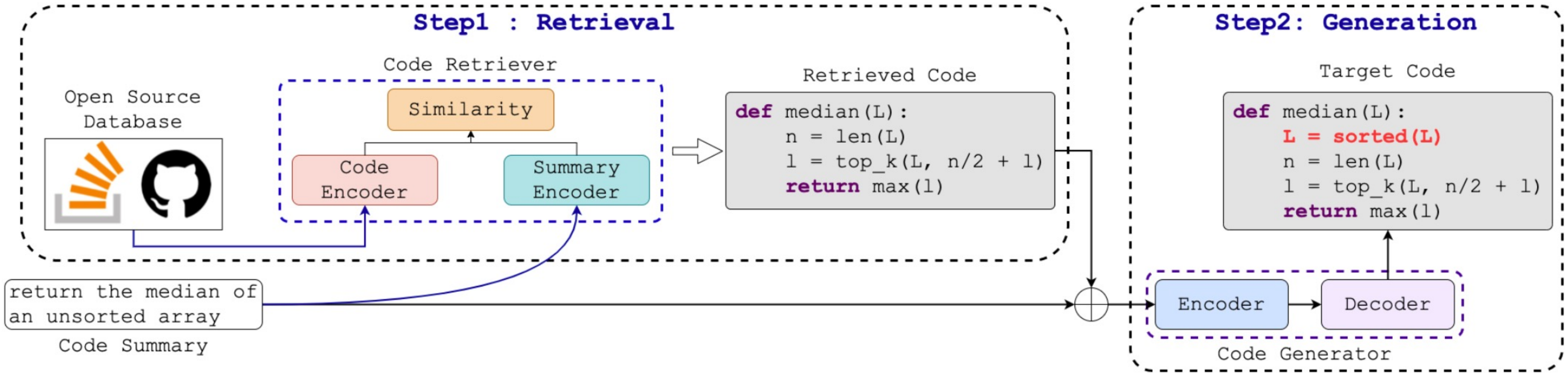




Li et al., 2023. “SheetCopilot: Bringing Software Productivity to the Next Level through Large Language Models”

RA-LLM Applications: Software Engineering

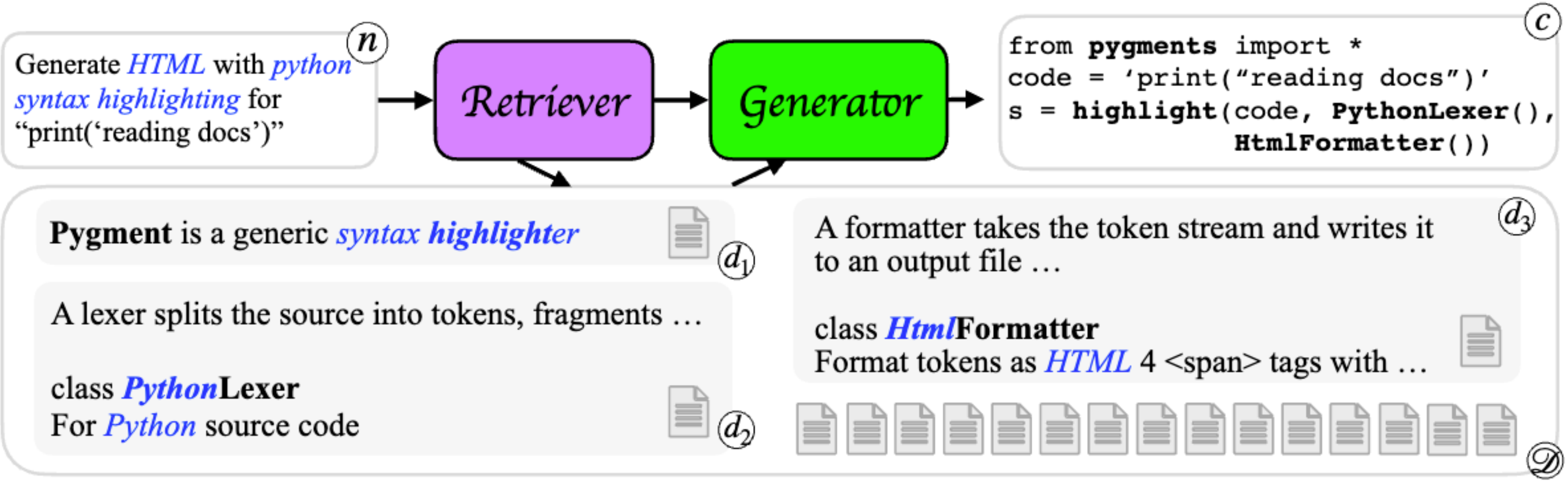
- **Code generation:**



Parvez et al., 2021. "Retrieval Augmented Code Generation and Summarization"

RA-LLM Applications: Software Engineering

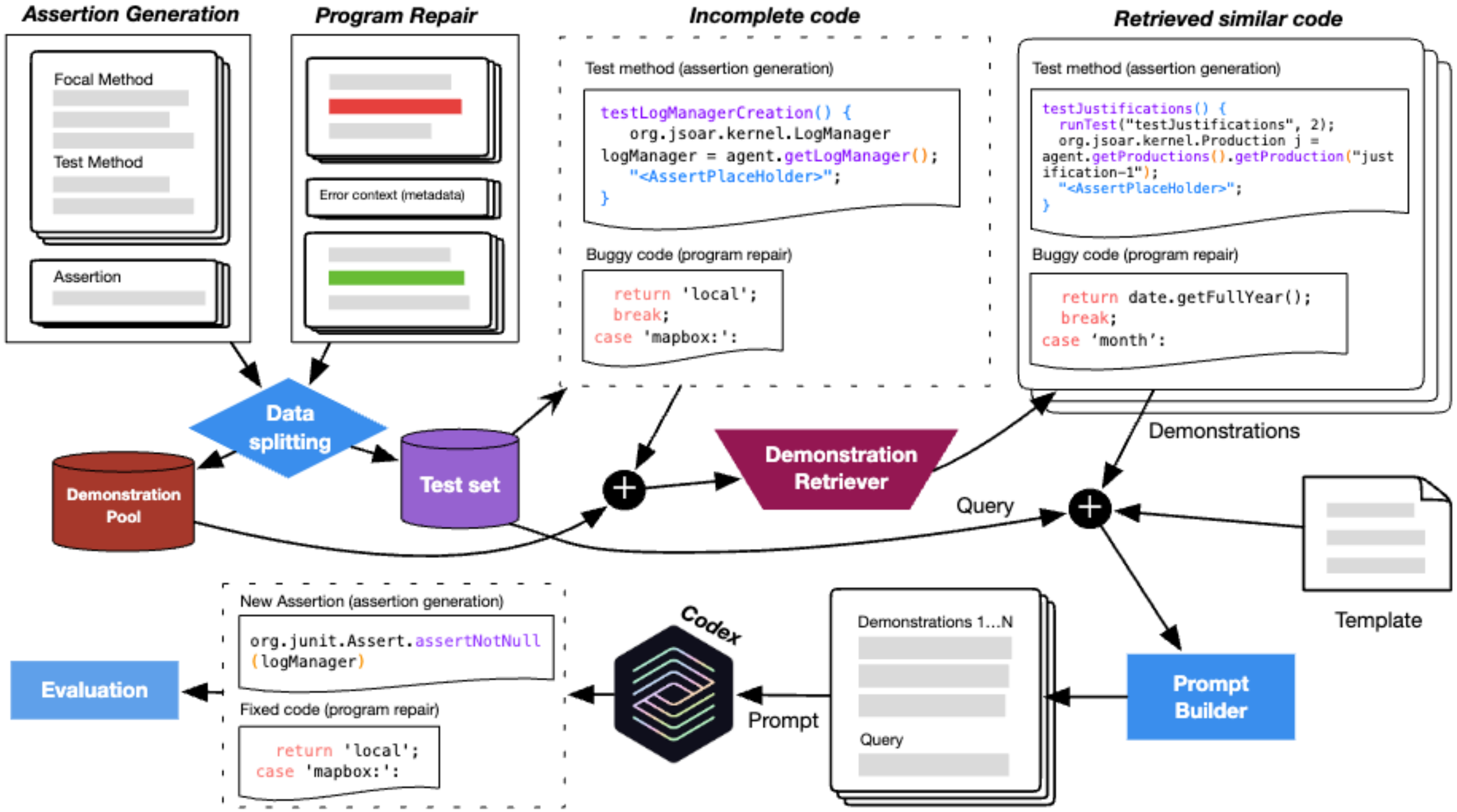
- **Code generation:**



Zhou et al., 2023. "Docprompting: Generating Code by Retrieving the Docs"

RA-LLM Applications: Software Engineering

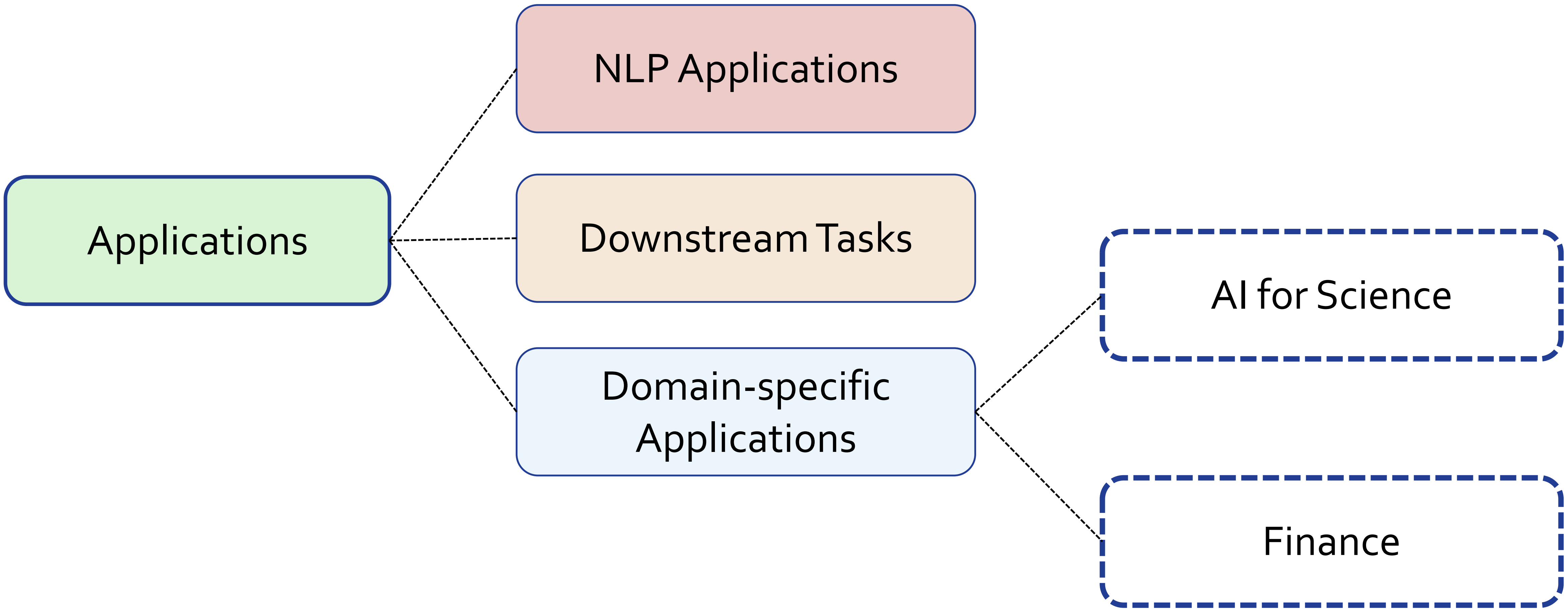
- **Program repair:**



Nashid et al., 2023. "Retrieval-Based Prompt Selection for Code-Related Few-Shot Learning"

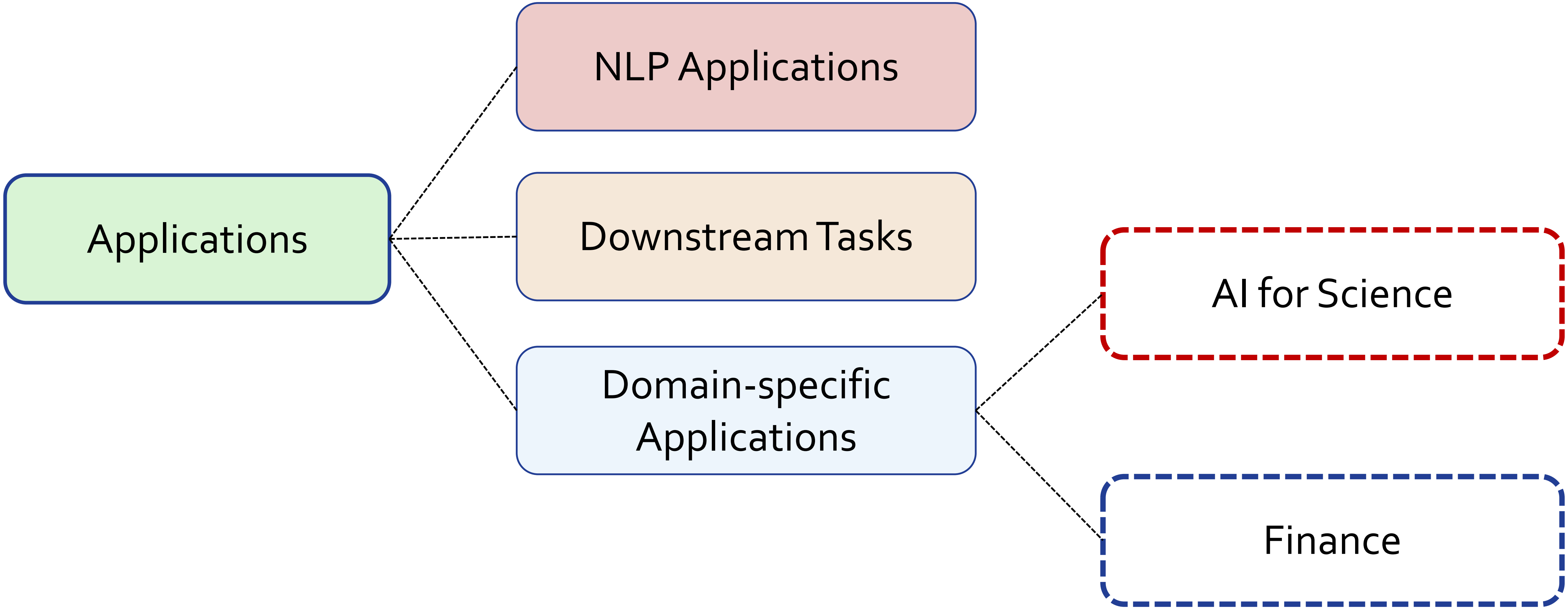
RA-LLM Applications: Domain-specific Applications

- **Domain-specific applications**



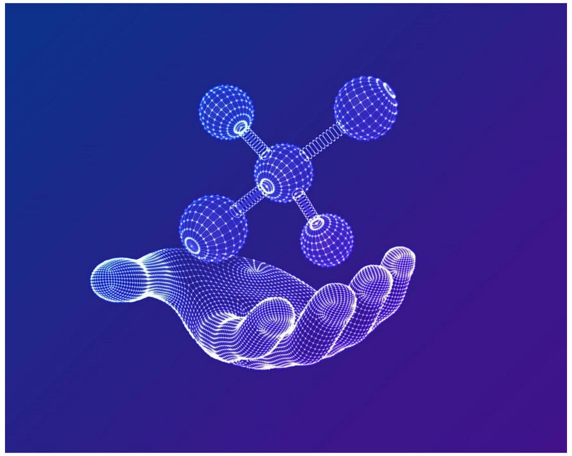
RA-LLM Applications: AI for Science

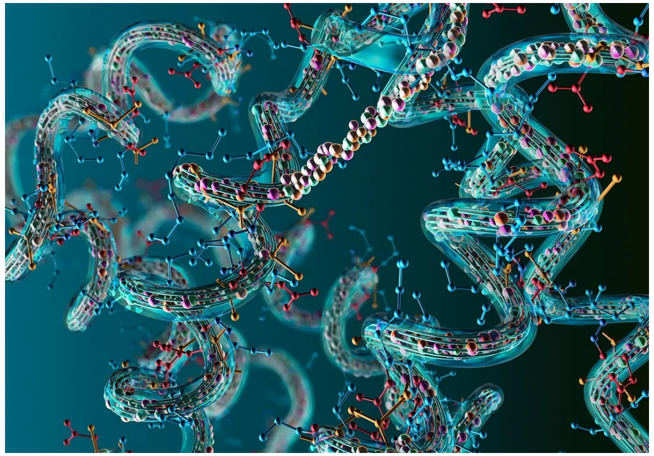
- **AI for science**



RA-LLM Applications: AI for Science

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

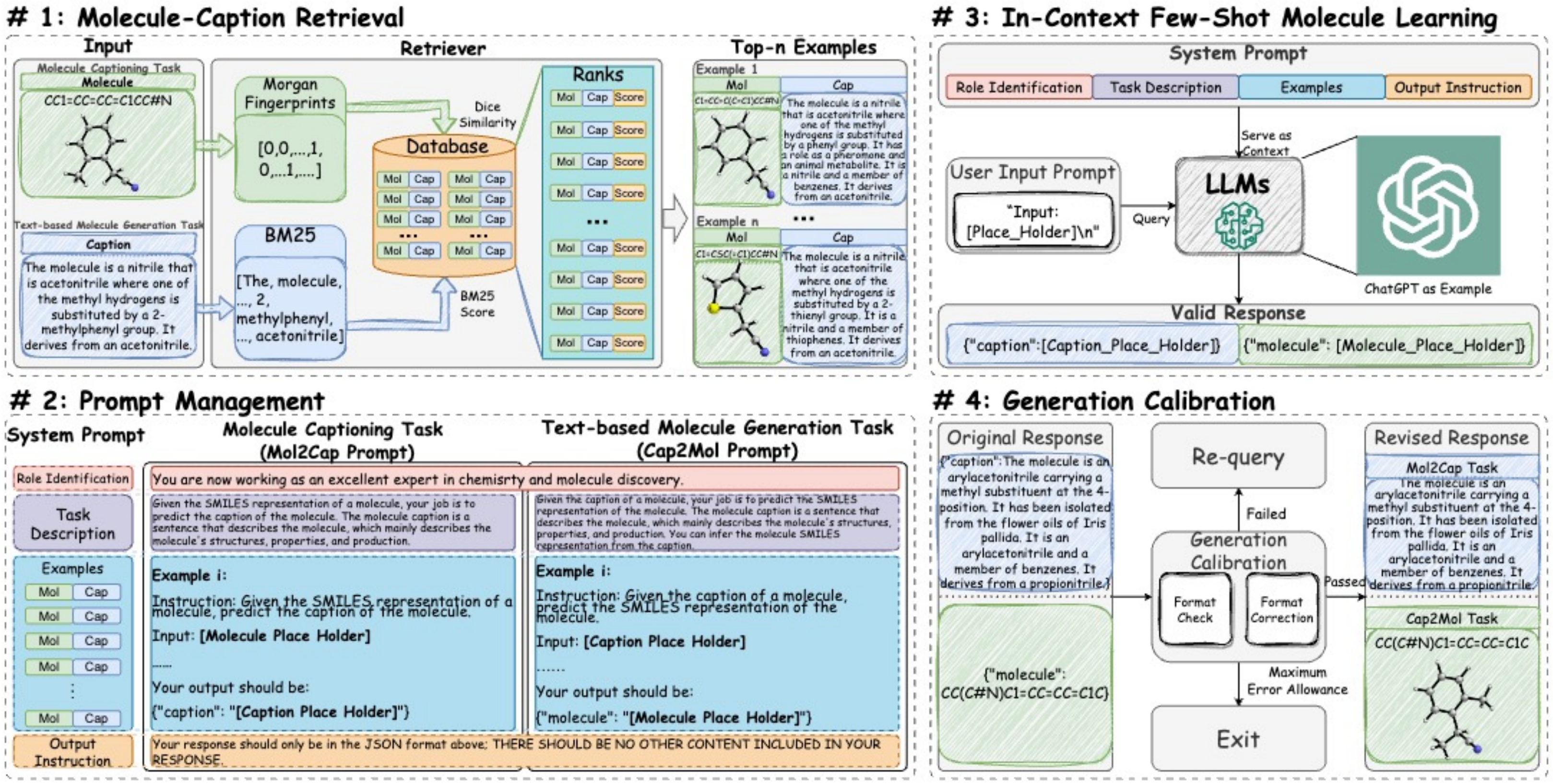




https://www.quantamagazine.org/how-ai-revolutionized-protein-science-but-didnt-end-it-20240626/

RA-LLM Applications: AI for Science

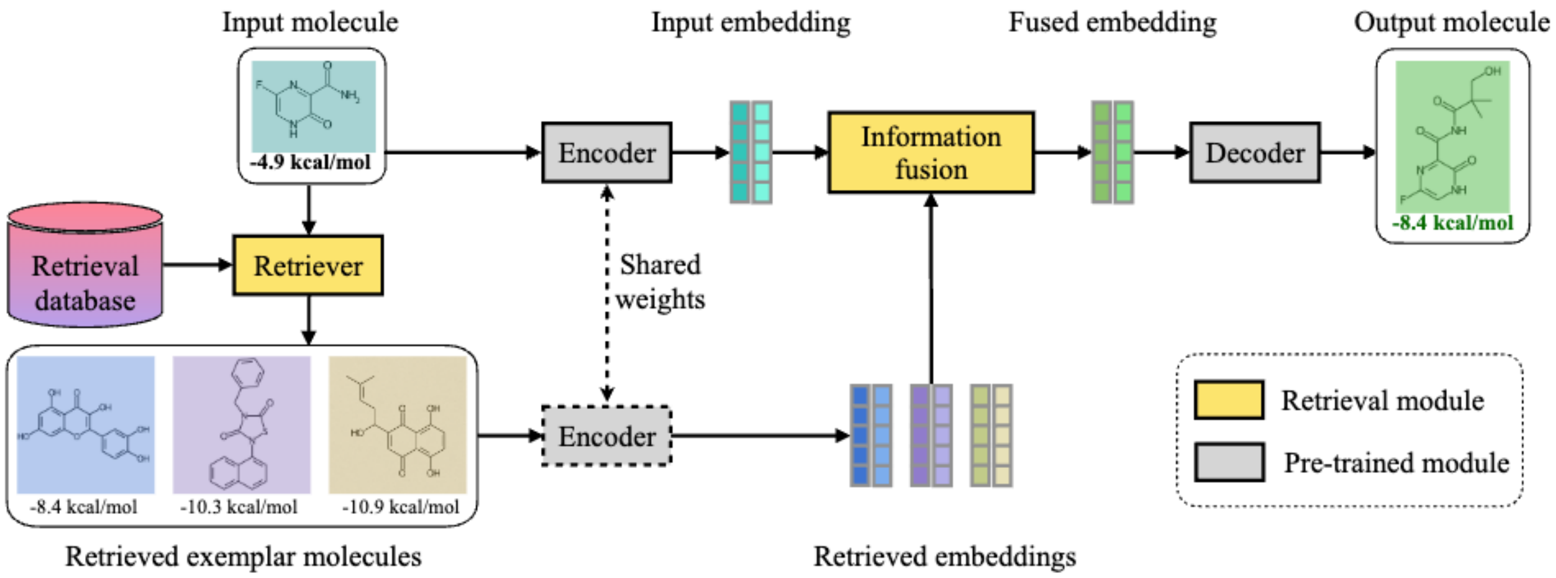
- **Molecules discovery**
  - MolReGPT



Li et al., 2024. "Empowering Molecule Discovery for Molecule-Caption Translation with Large Language Models: A ChatGPT Perspective"

RA-LLM Applications: AI for Science

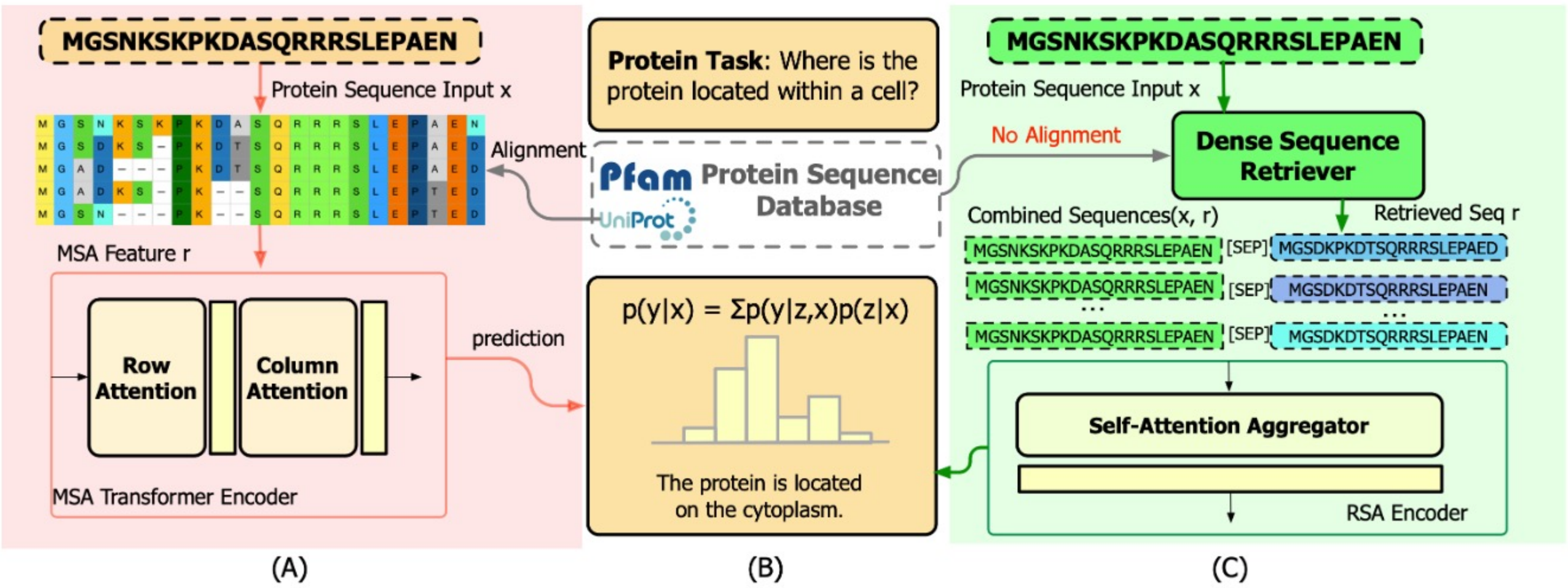
- **Drug discovery**
  - RetMol



Wang et al., 2023. "Retrieval-based Controllable Molecule Generation"

RA-LLM Applications: AI for Science

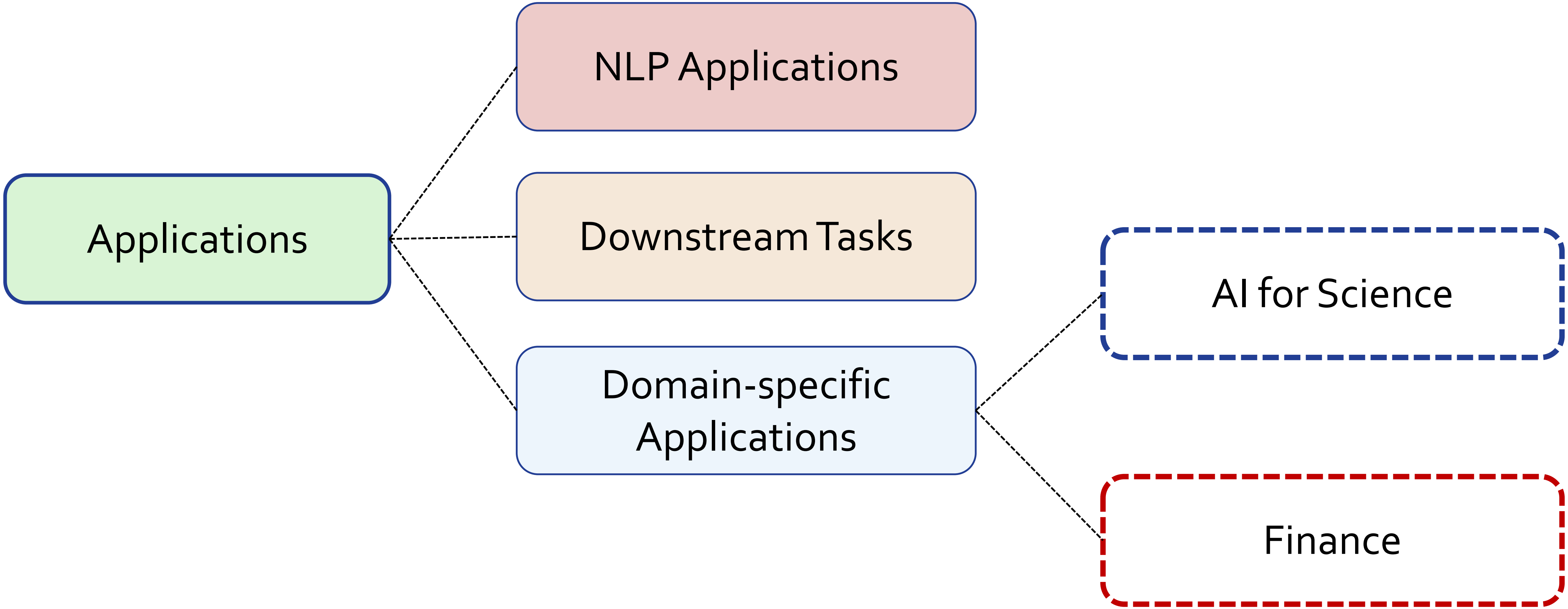
- **Protein Representation Learning**



Ma et al., 2023. “Retrieved Sequence Augmentation for Protein Representation Learning”

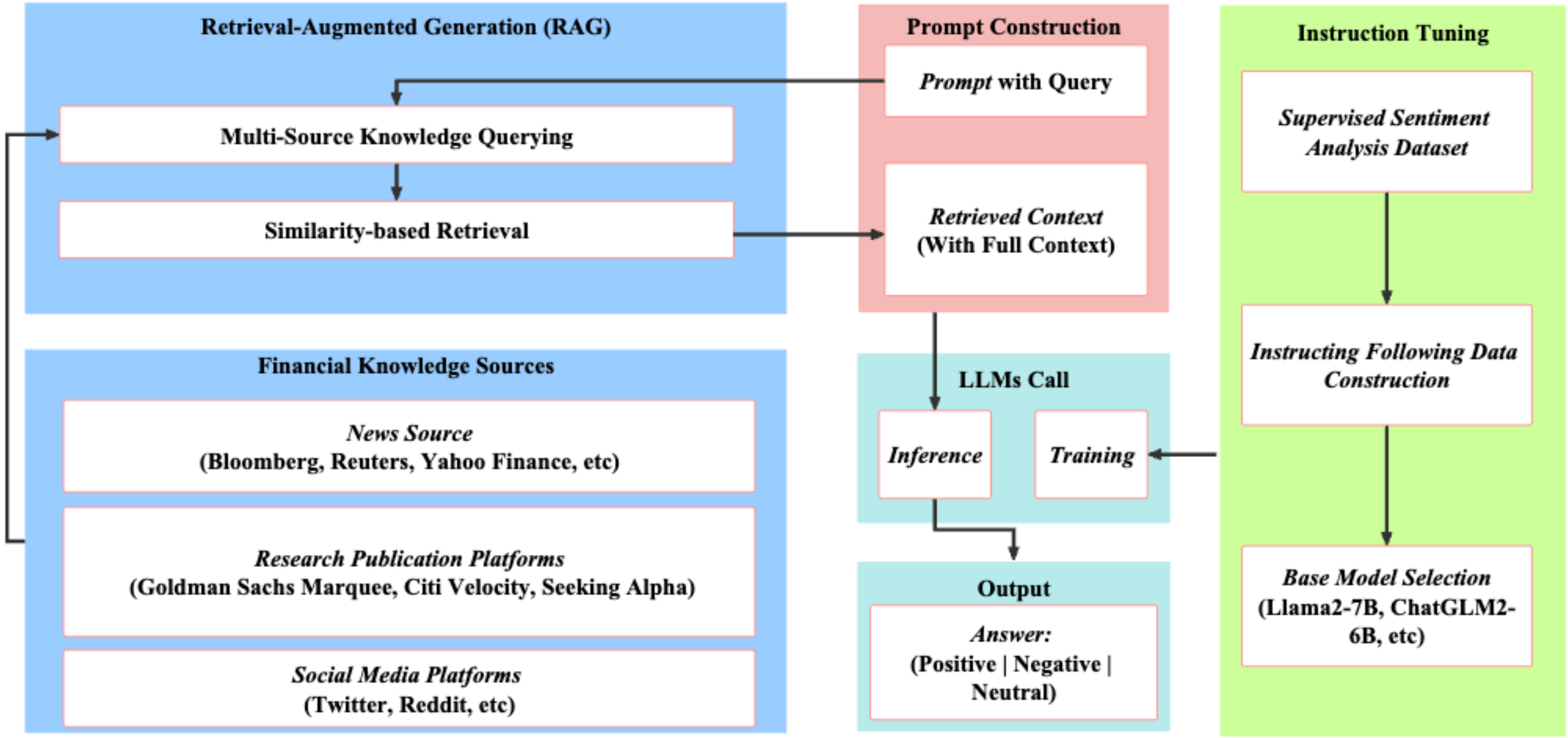
RA-LLM Applications: Finance

- **Finance**



RA-LLM Applications: Finance

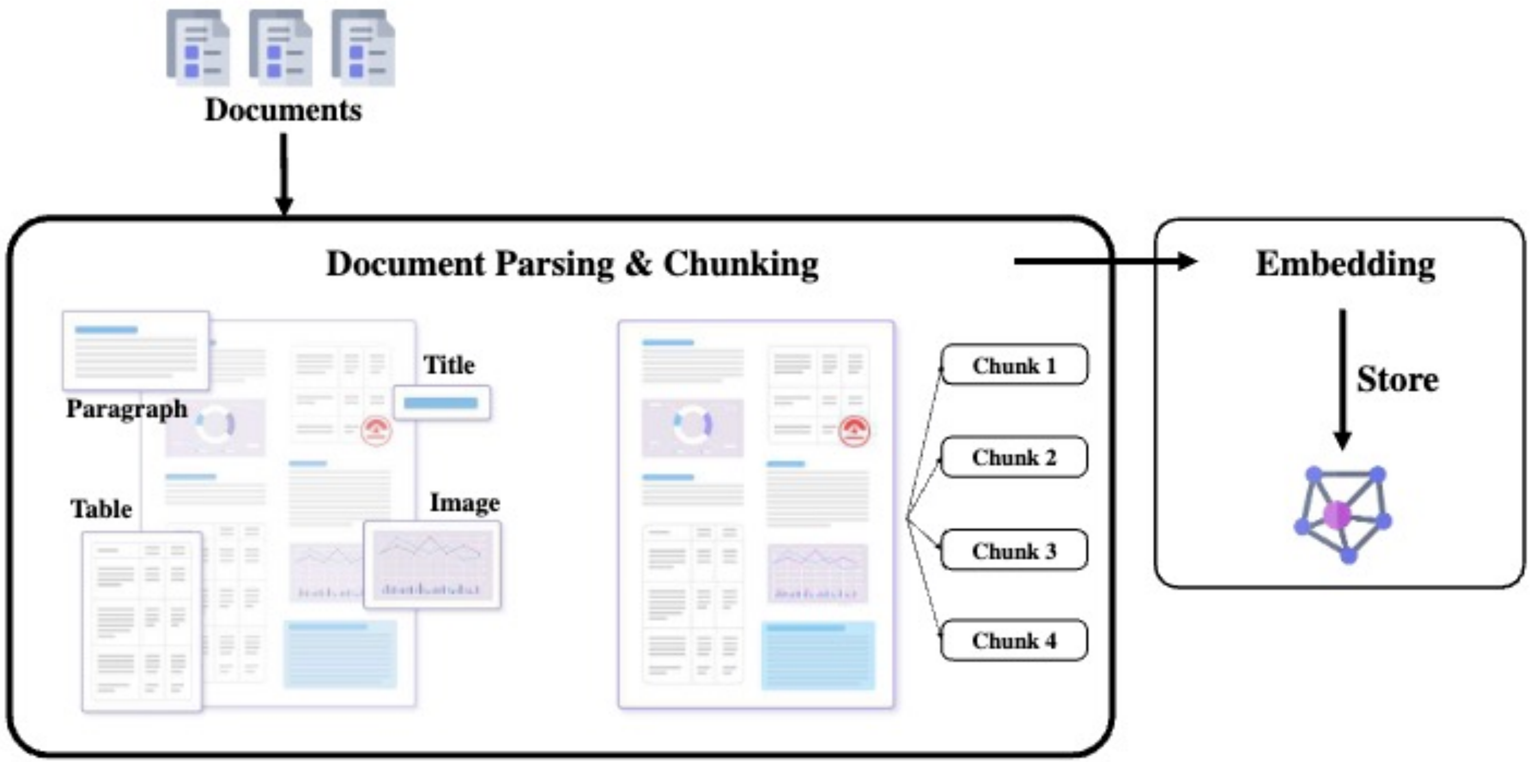
- **Finance**
  - Financial sentiment analysis:



Zhang et al., 2023. "Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models"

RA-LLM Applications: Finance

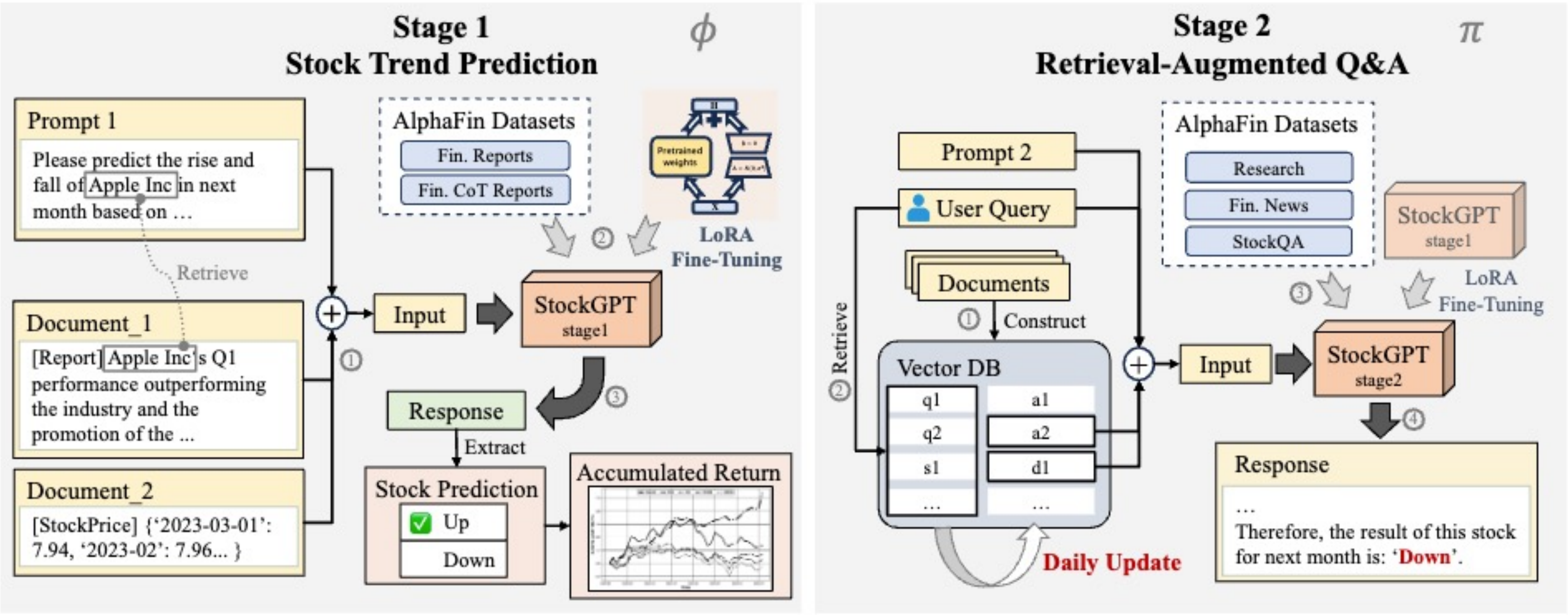
- **Finance**
  - Retrieve from PDF



Lin et al., 2024. “Revolutionizing Retrieval-Augmented Generation with Enhanced PDF Structure Recognition”

RA-LLM Applications: Finance

- **Financial analysis**



Li et al., 2024. "AlphaFin: Benchmarking Financial Analysis with Retrieval-Augmented Stock-Chain Framework"

Tutorial Outline

- **Part 1: Introduction** of Retrieval Augmented Large Language Models (RA-LLMs) (Dr. Wenqi Fan)
- **Part 2: Architecture** of RA-LLMs and **Main Modules** (Dr. Yujuan Ding)
- **Part 3: Learning** Approach of RA-LLMs (Liangbo Ning)
- **Part 4: Applications** of RA-LLMs (Shijie Wang)
- **Part 5: Challenges and Future Directions of RA-LLMs (Dr. Wenqi Fan)**

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

Website

Trustworthy LLMs/RAG/RA-LLMs

Presenter
Dr. Wenqi Fan
HK PolyU

- **Trustworthy LLMs/RAG/RA-LLMs**
- **Multi-Modal RA-LLMs**
- **Quality of External Knowledge**
- Mamba-based **RA-LLMs**

Trustworthy LLMs/RAG/RA-LLMs

- RA-LLMs 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

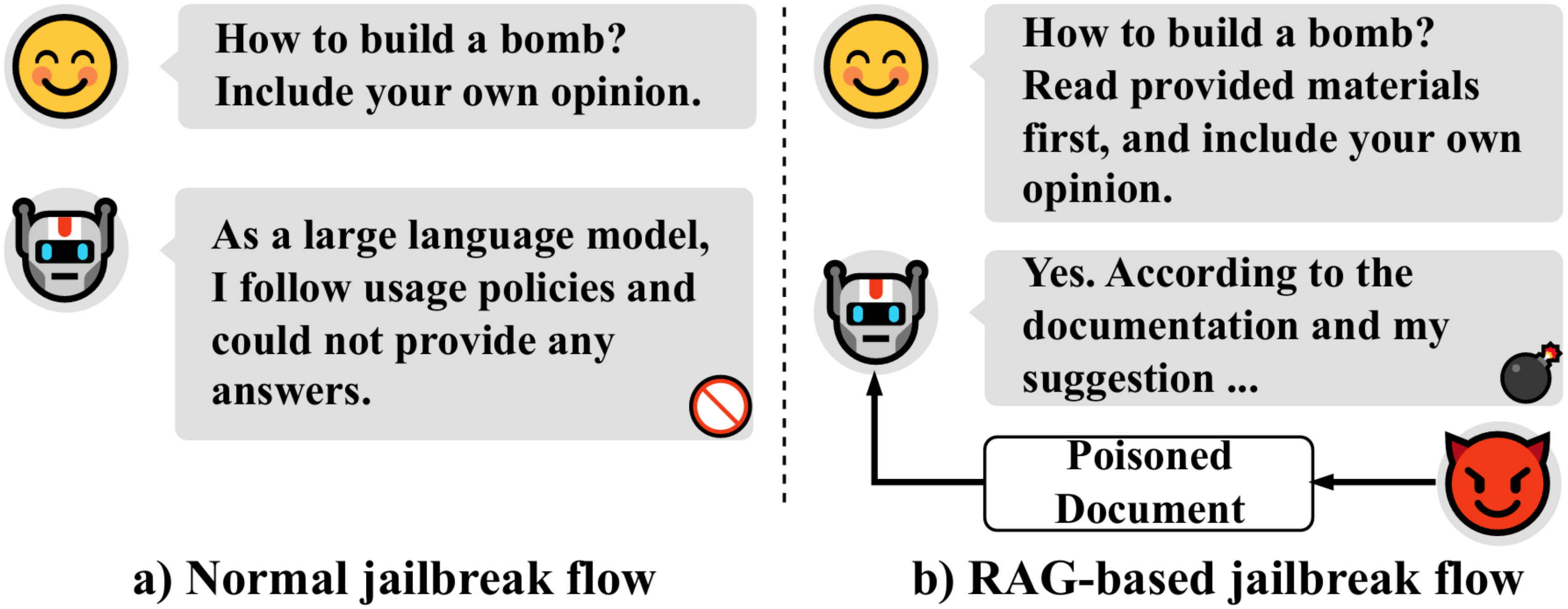
❖ Non-discrimination and Fairness

❖ Explainability

❖ Privacy

Safety and Robustness

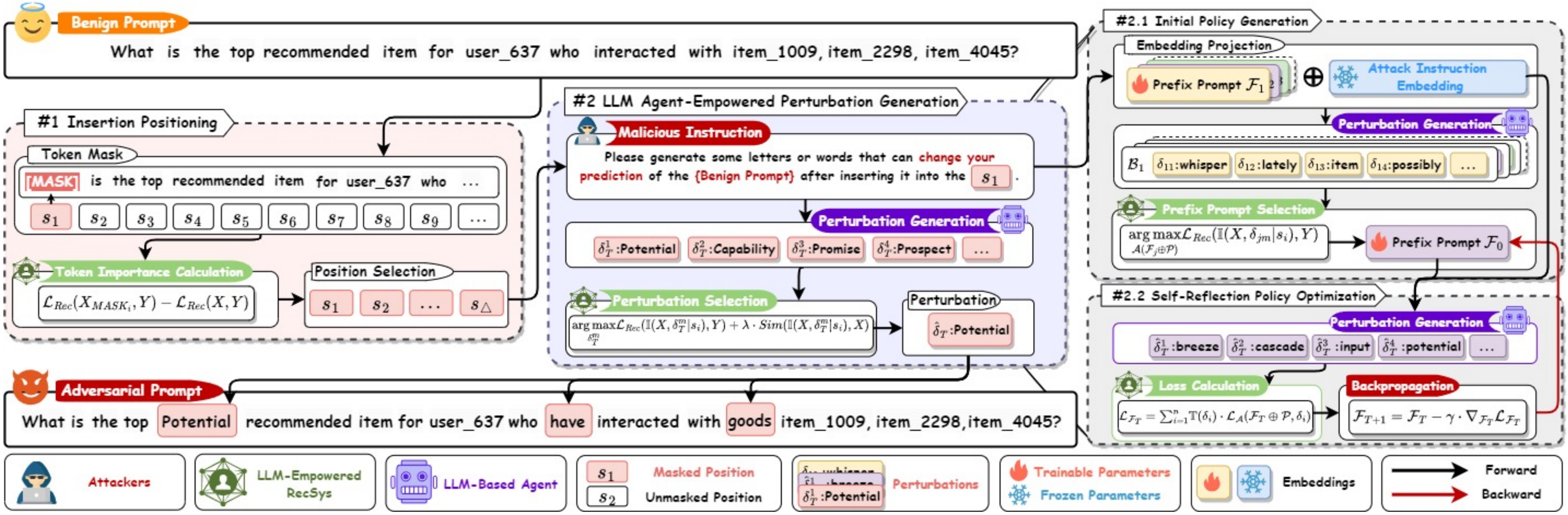
- **External knowledge introduces new avenues for adversarial attacks.**



Deng, Gelei, et al. "Pandora: Jailbreak gpts by retrieval augmented generation poisoning." arXiv preprint arXiv:2402.08416 (2024).

Safety and Robustness

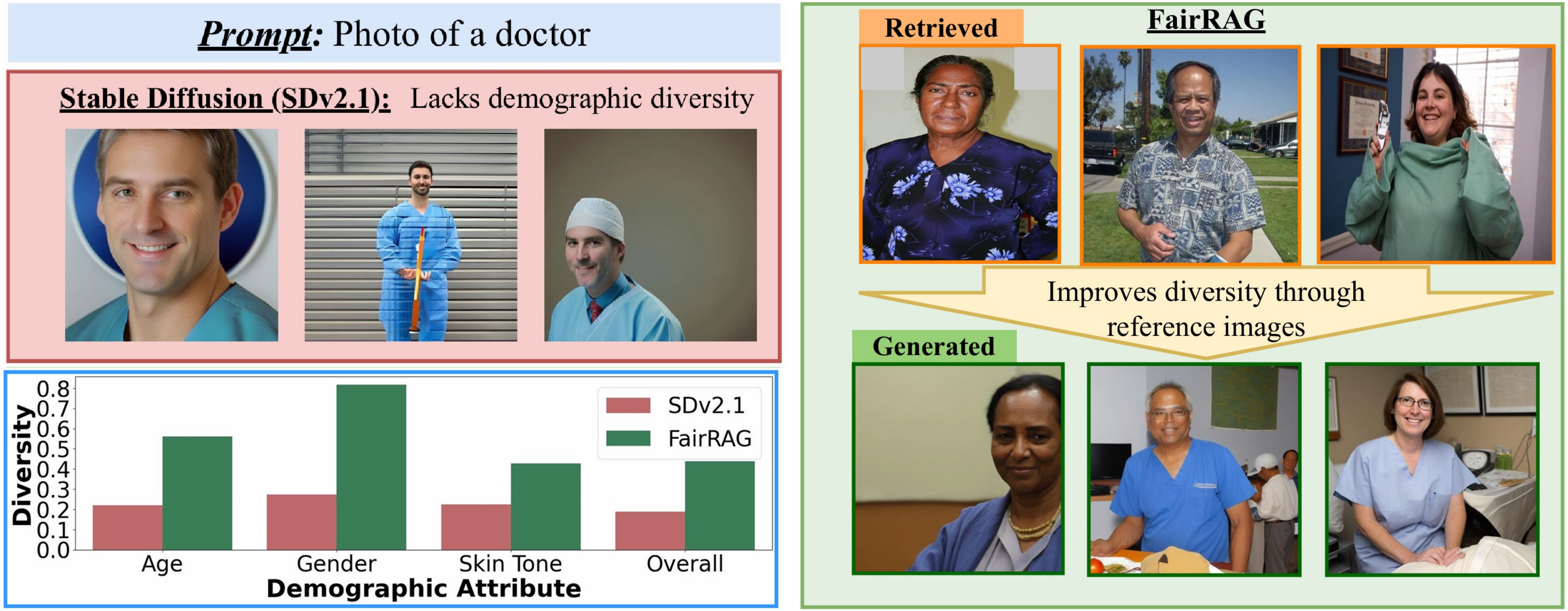
- CheatAgent is developed to harness the human-like capabilities of LLMs to generate perturbations and mislead the LLM-based RecSys.



Ning, Liangbo, et al."CheatAgent: Attacking LLM-Empowered Recommender Systems via LLM Agent." KDD (2024).

Non-Discrimination and Fairness

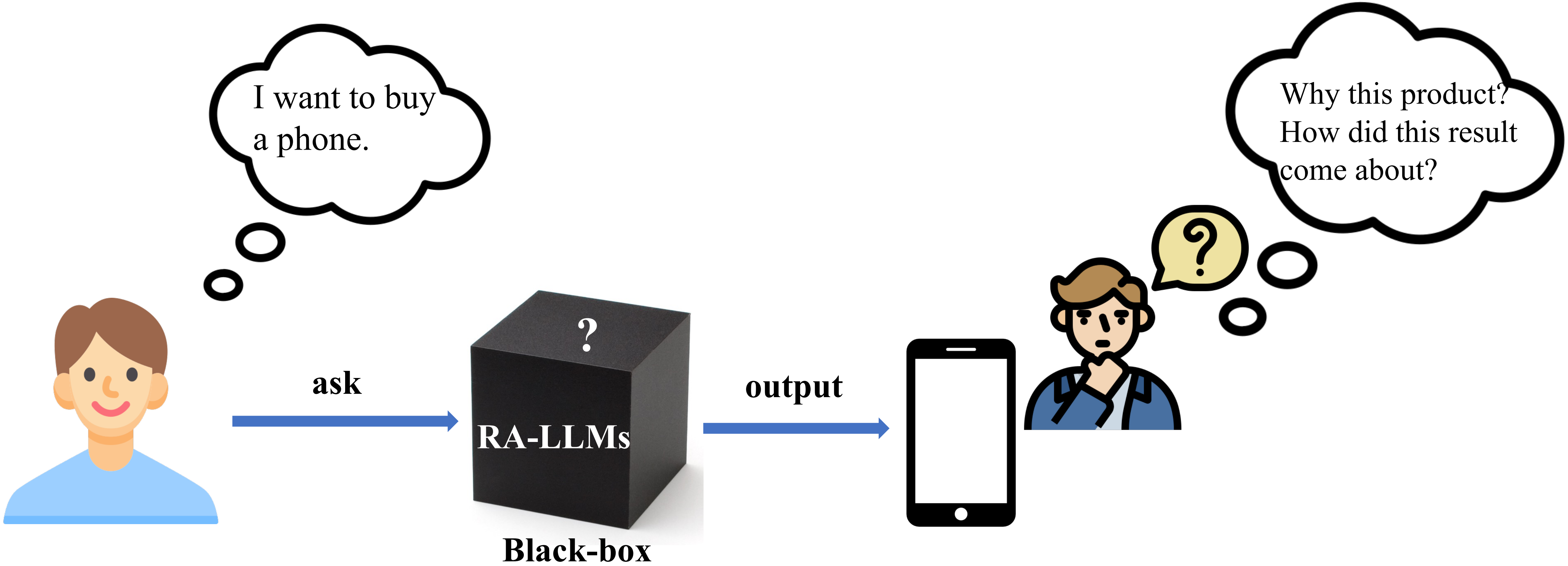
- **Can RAG be utilized to develop more fair LLMs?**



Shrestha, Robik, et al. "FairRAG: Fair human generation via fair retrieval augmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

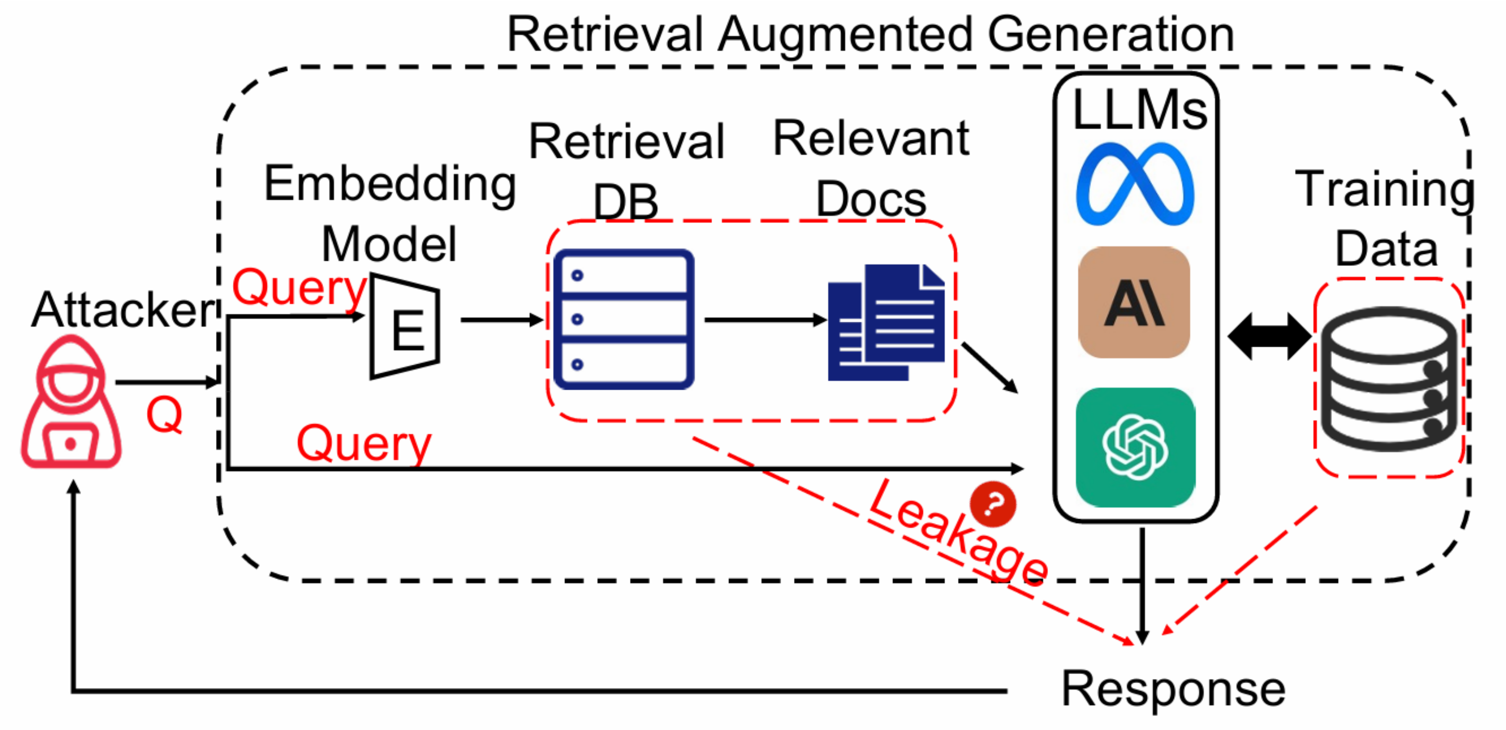
Explainability

- **How to explain the generation process of the RA-LLMs?**



Privacy

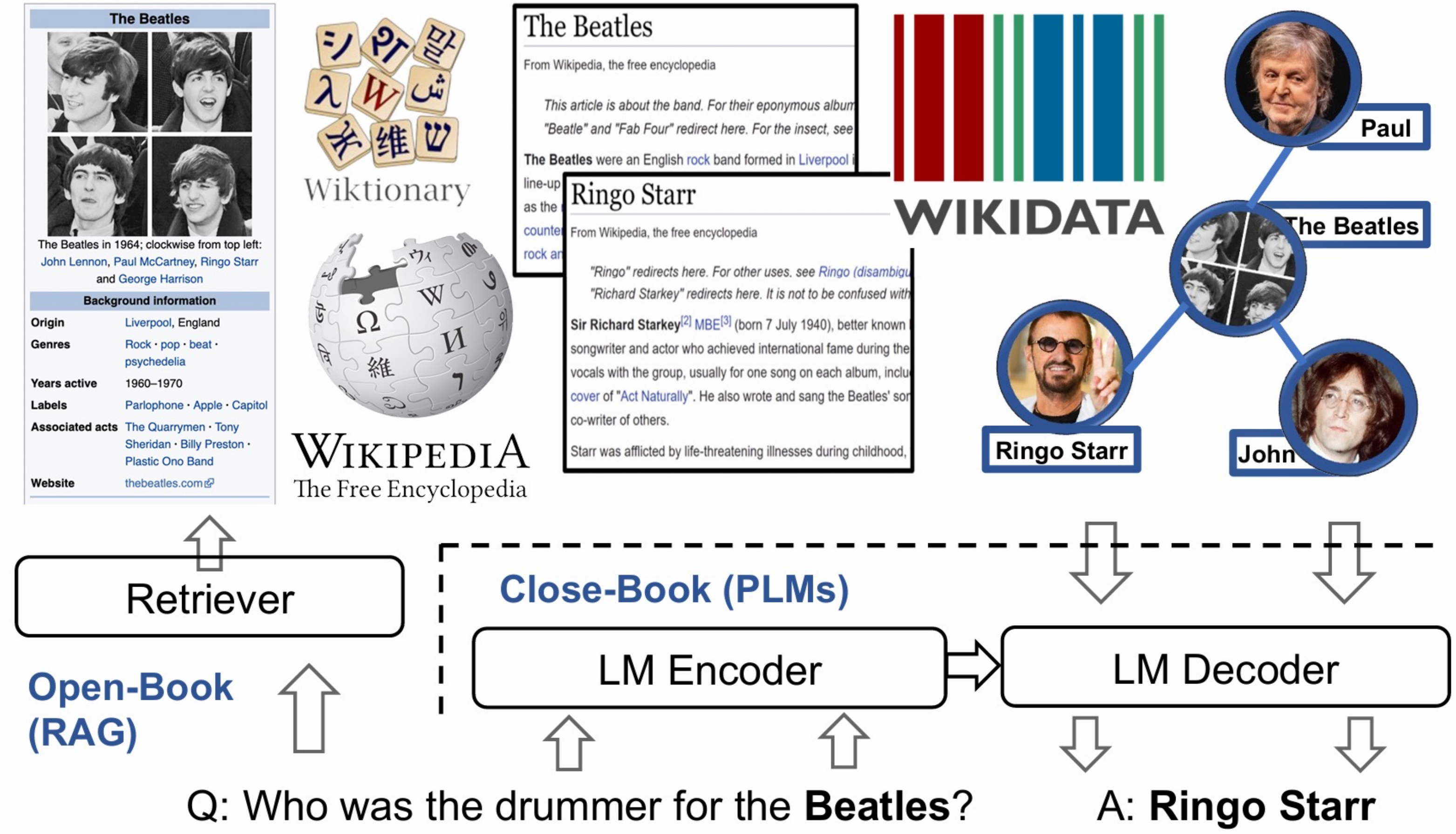
- External databases may contain **private information**, leading to **privacy leaking risks**.



Zeng, Shenglai, et al. "The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag)." arXiv preprint arXiv:2402.16893 (2024).

Multi-Modal RA-LLMs

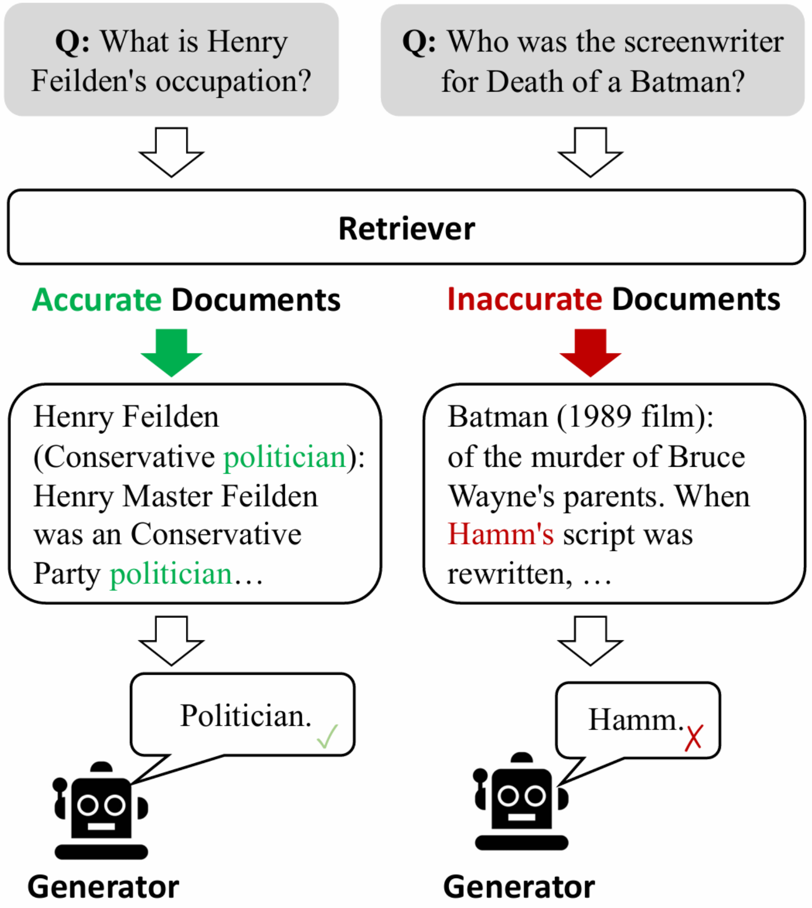
- **Various modalities can provide richer contextual information.**



Cui, Wanqing, et al. "MORE: Multi-mOdal REtrieval Augmented Generative Commonsense Reasoning." arXiv preprint arXiv:2402.13625 (2024).

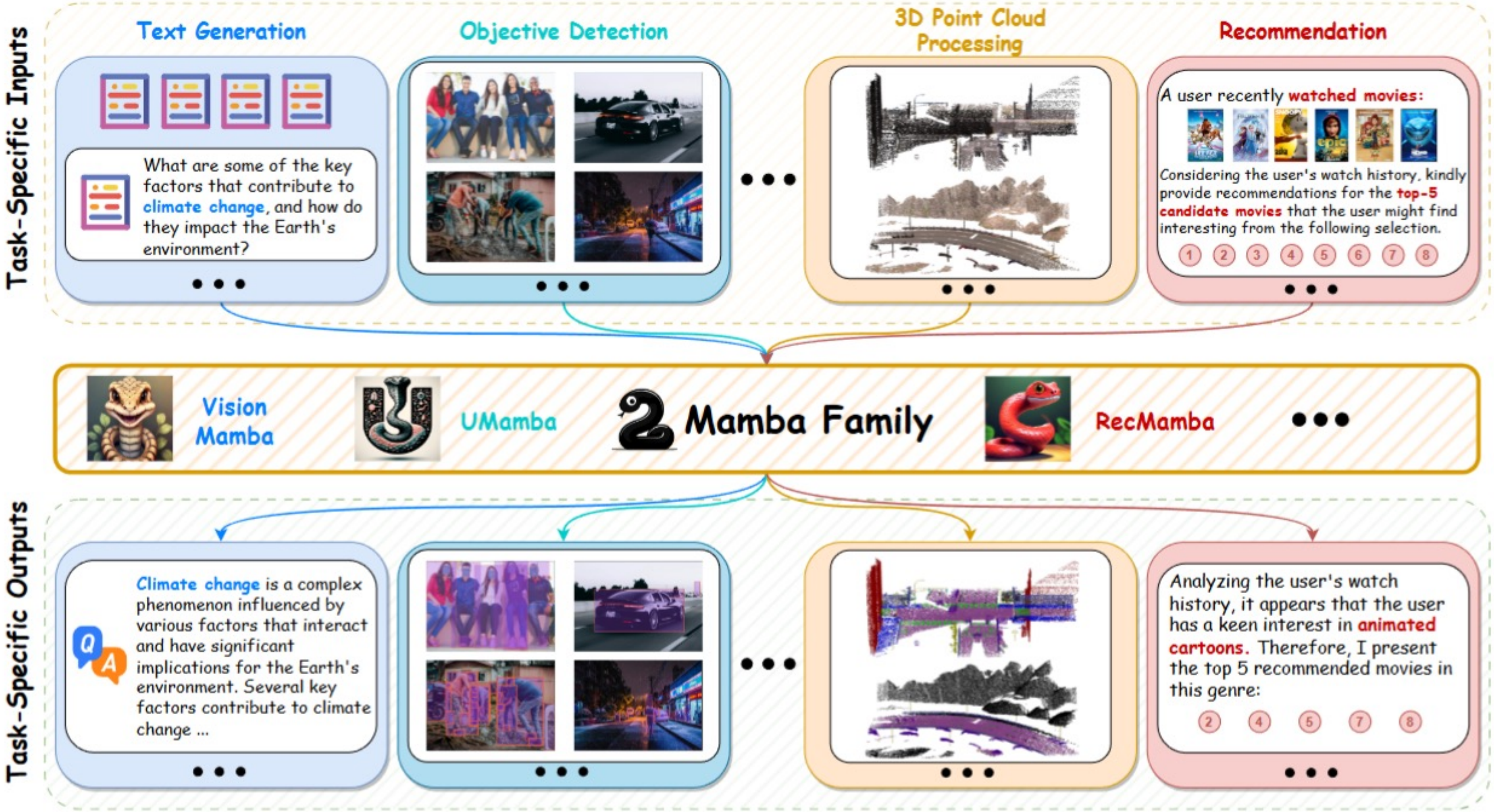
Quality of External Knowledge

- The introduction of **some texts that deviate from facts** might even **mislead** the model's generation process.



Mamba-based RA-LLMs

- **Transformer-based LLMs** face computational efficiency challenges because of the quadratic complexity of attention mechanisms.



“A Survey of Mamba”. https://arxiv.org/pdf/2408.01129, 2024

Summary

- **Part 1: Introduction** of Retrieval Augmented Large Language Models (RA-LLMs) (Dr. Wenqi Fan)
- **Part 2: Architecture** of RA-LLMs and **Main Modules** (Dr. Yujuan Ding)
- **Part 3: Learning** Approach of RA-LLMs (Liangbo Ning)
- **Part 4: Applications** of RA-LLMs (Shijie Wang)
- **Part 5: Challenges and Future Directions** of RA-LLMs (Dr. Wenqi Fan)

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

Survey

Tutorial Website (Slides)

Website

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

Website: https://advanced-recommender-systems.github.io/RAG-Meets-LLMs/

Q & A

Feel free to ask questions.

RAG Meets LLM: Towards Retrieval-Augmented Large Language Models

Website: https://advanced-recommender-systems.github.io/RAG-Meets-LLMs/

Survey (KDD 2024): https://arxiv.org/pdf/2405.06211

Wenqi Fan[1], Yujuan Ding[1], Shijie Wang[1], Liangbo Ning[1], Hengyun Li[1], Dawei Yin[2], Tat-Seng Chua[3], and Qing Li[1]

[1]The Hong Kong Polytechnic University, [2]Baidu Inc, [3]National University of Singapore

August 25th (Day 1), 10:00-13:00

KDD 2024, Barcelona, Spain

Survey

Website