SIGIR2026
20-24 July
MELBOURNE | NAARM
AUSTRALIA

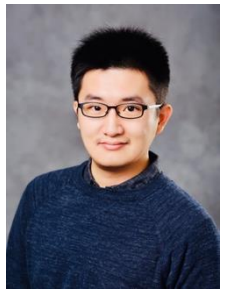
Bridging Personalization and AI: From RAG to Agent











Pengyue Jia[1] Xiaopeng Li[1] Derong Xu[1] Yi Wen[1] Yingyi Zhang[1] Wenlin Zhang[1] Yichao Wang[2] Yong Liu[2] Xiangyu Zhao[1]

[1]City University of Hong Kong, [2]Huawei Noah's Ark Lab

CityU AML Lab

Huawei Noah's Ark Lab

Agenda

1. Introduction
9:00 – 9:50



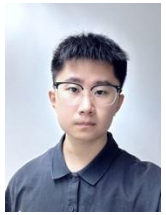
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



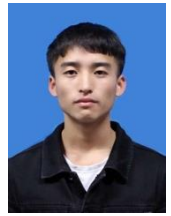
Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



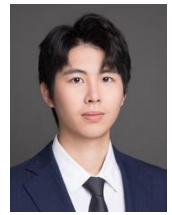
Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

8. Hands on Projects
16:45 – 17:30



Pengyue Jia

Agenda

1. Introduction
9:00 – 9:50



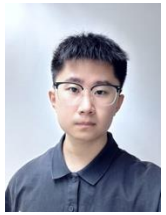
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

8. Hands on Projects
16:45 – 17:30



Pengyue Jia

Why We Need Personalization?

Age of Information Explosion

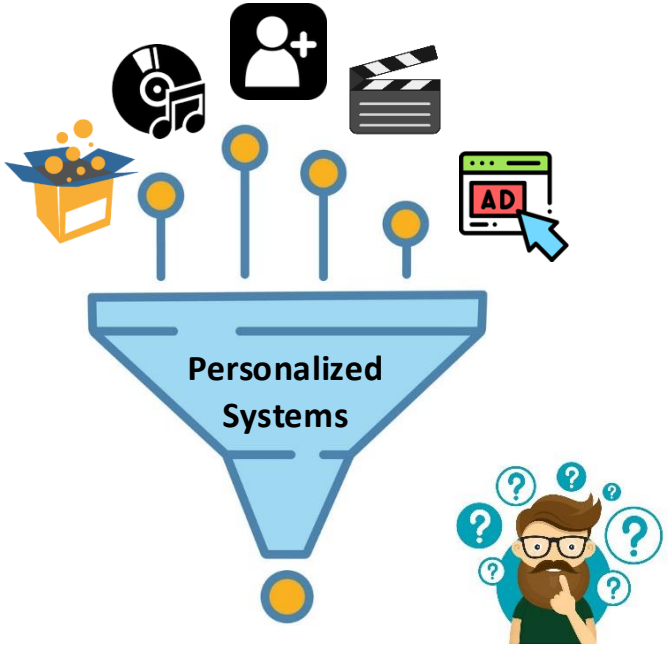


Why We Need Personalization?

Age of Information Explosion



Personalized Service



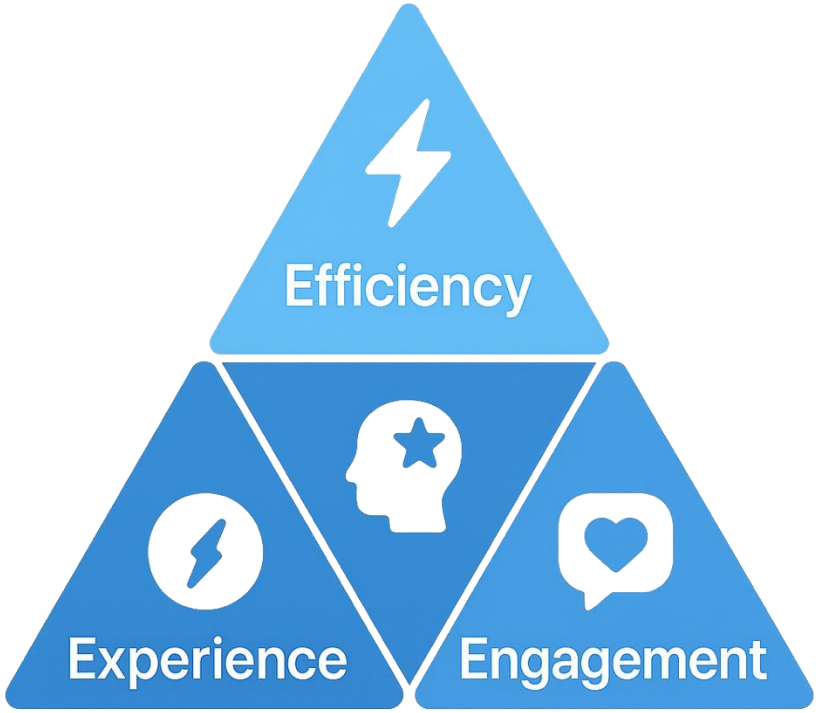
Application Scenarios

Product Interaction, AI Assistant , Customer Service, IoT environment

Personalization

Boost Efficiency

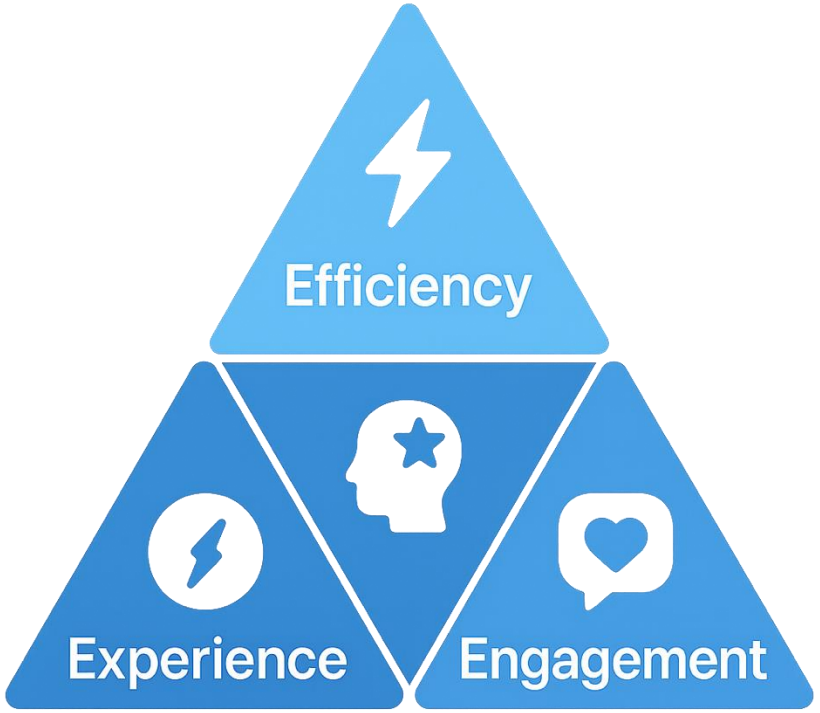
Lower cognitive & interaction cost



Personalization

Boost Efficiency

Lower cognitive & interaction cost



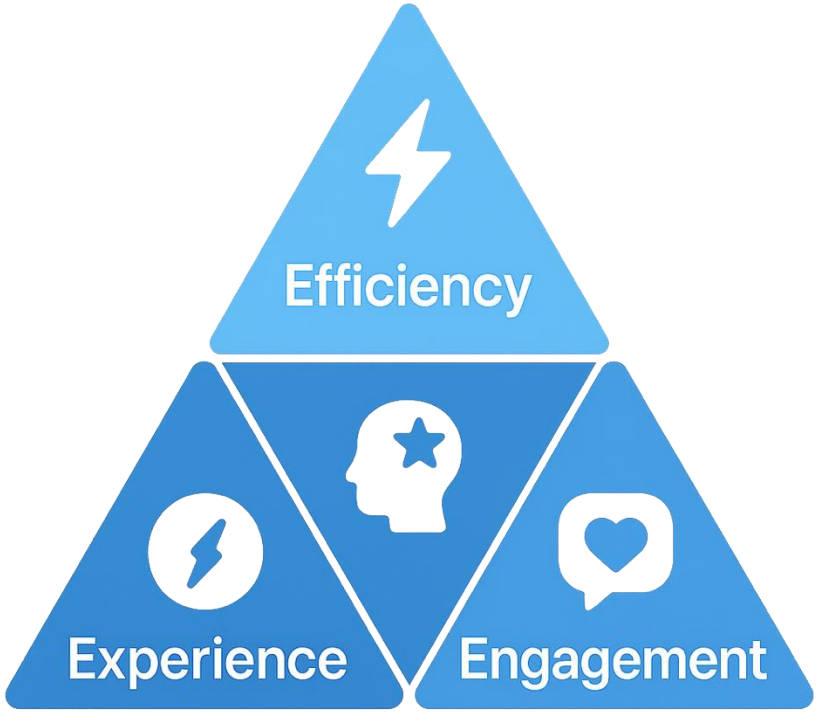
Increase Engagement

Co-create value & raise LTV

Personalization

Boost Efficiency

Lower cognitive & interaction cost



Elevate Experience

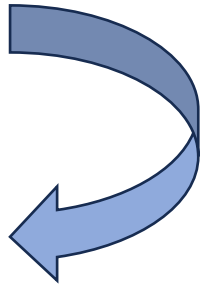
More intuitive & delightful UX

Increase Engagement

Co-create value & raise LTV

Personalization Definition

Personalization is tailoring information to an individual user's specific needs and preferences.



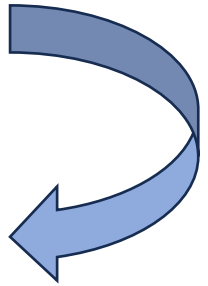
Personalization **adapts RAG and agent systems to individual users** by leveraging their preferences, contexts, histories, and goals **across retrieval, reasoning, generation, and actions**.

- Adapt system behavior to individual users
- Capture preferences, contexts, histories, and goals
- Transform generic outputs into tailored experiences
- In RAG/Agent: personalize retrieval, reasoning, generation, and actions

[1] Tiihonen, Juha, and Alexander Felfernig. "An introduction to personalization and mass customization." Journal of intelligent information systems 49.1 (2017): 1-7.

Personalization Definition

Personalization is tailoring information to an individual user's specific needs and preferences.



Personalization **adapts RAG and agent systems to individual users** by leveraging their preferences, contexts, histories, and goals **across retrieval, reasoning, generation, and actions**.

- Adapt system behavior to individual users
- Capture preferences, contexts, histories, and goals
- Transform generic outputs into tailored experiences
- In RAG/Agent: personalize retrieval, reasoning, generation, and actions

[1] Tiihonen, Juha, and Alexander Felfernig. "An introduction to personalization and mass customization." Journal of intelligent information systems 49.1 (2017): 1-7.

Traditional Personalization Method

Rule-based Method



Recommender System

Traditional Personalization Method

Rule-based Method → **Collaborative Filtering/ Matrix Factorization**



Recommender System



Search Engine

Traditional Personalization Method

Rule-based Method → **Collaborative Filtering/ Matrix Factorization** → **Feature Engineering + Machine Learning Model**



Recommender System



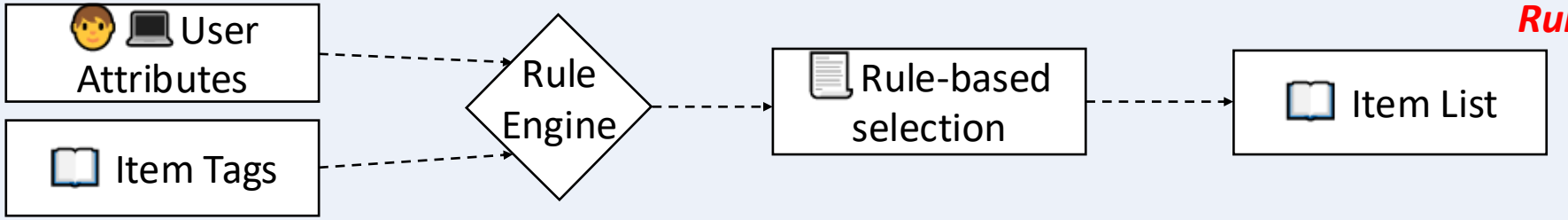
Search Engine



Online Advertising

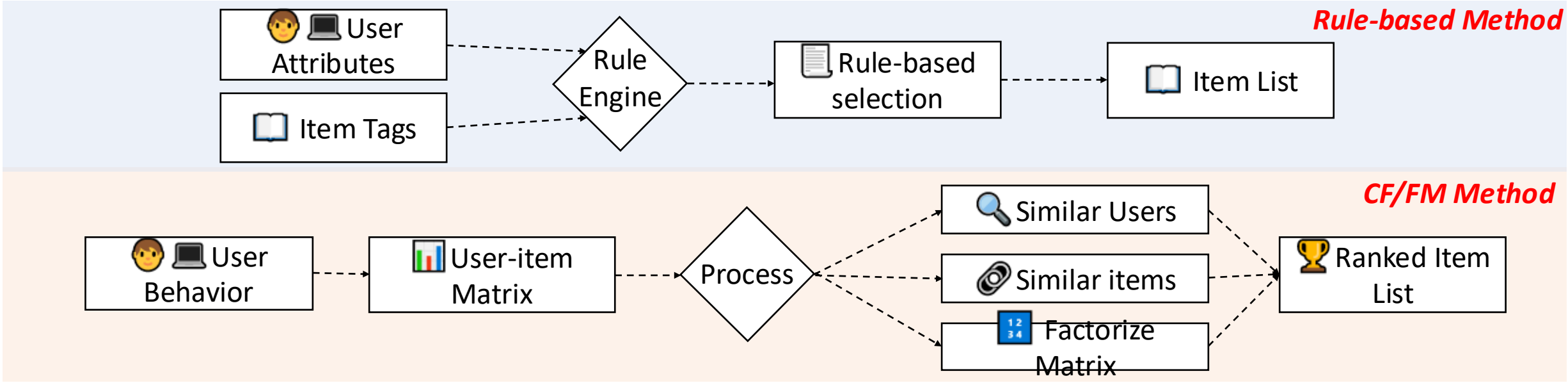
Traditional Personalization Method

➢Case: Recommend Shopping Items for Users



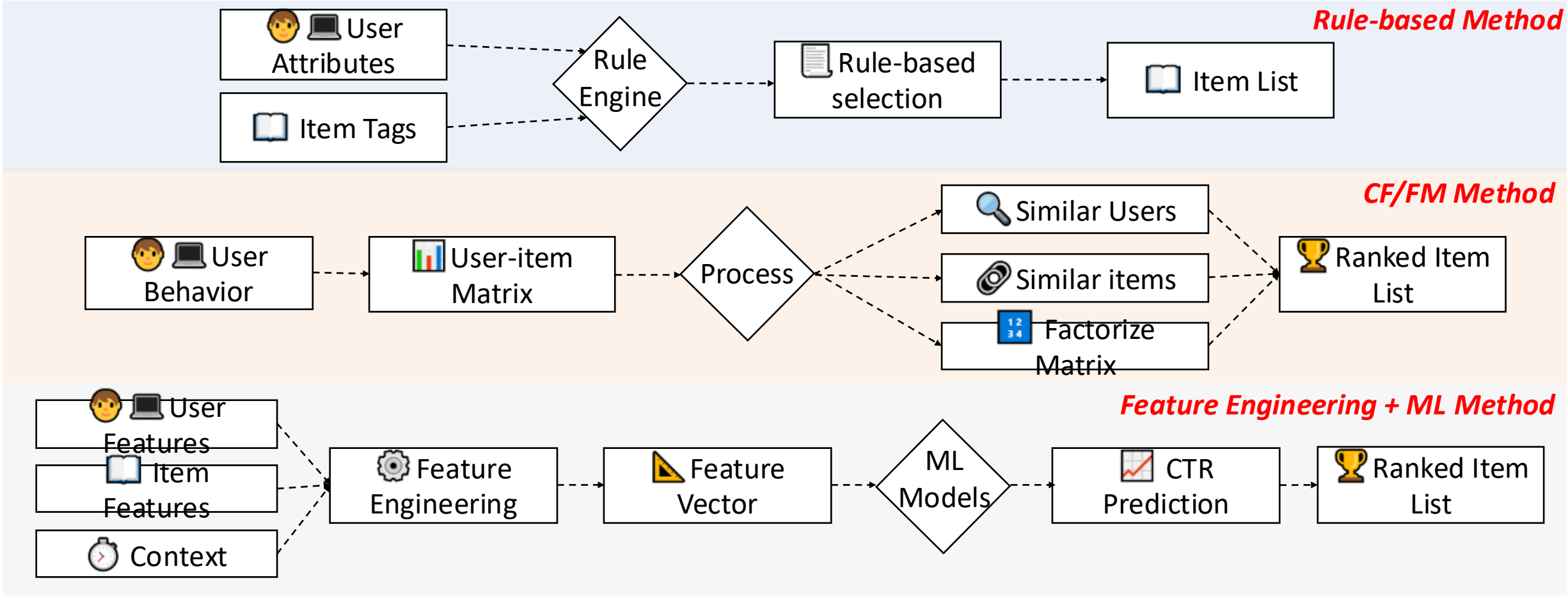
Traditional Personalization Method

➢Case: Recommend Shopping Items for Users



Traditional Personalization Method

➢Case: Recommend Shopping Items for Users



Traditional Personalization Method Limitation

Data Silos



Database A **Database B**

- Data is scattered on different systems
- Bring disconnected service for users

Traditional Personalization Method Limitation

Data Silos



Database A Database B

- Data is scattered on different systems
- Bring disconnected service for users

Insufficient User Understanding



- Unable to capture authentic and real-time user intent
- Might provide un-relevant content to users

Traditional Personalization Method Limitation

Data Silos



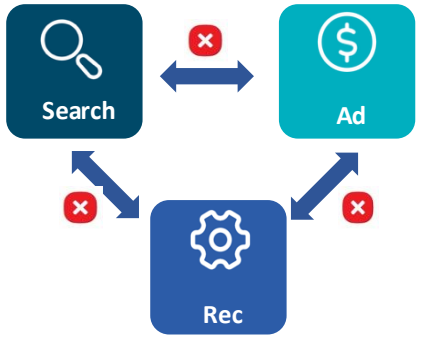
- Data is scattered on different systems
- Bring disconnected service for users

Insufficient User Understanding



- Unable to capture authentic and real-time user intent
- Might provide un-relevant content to users

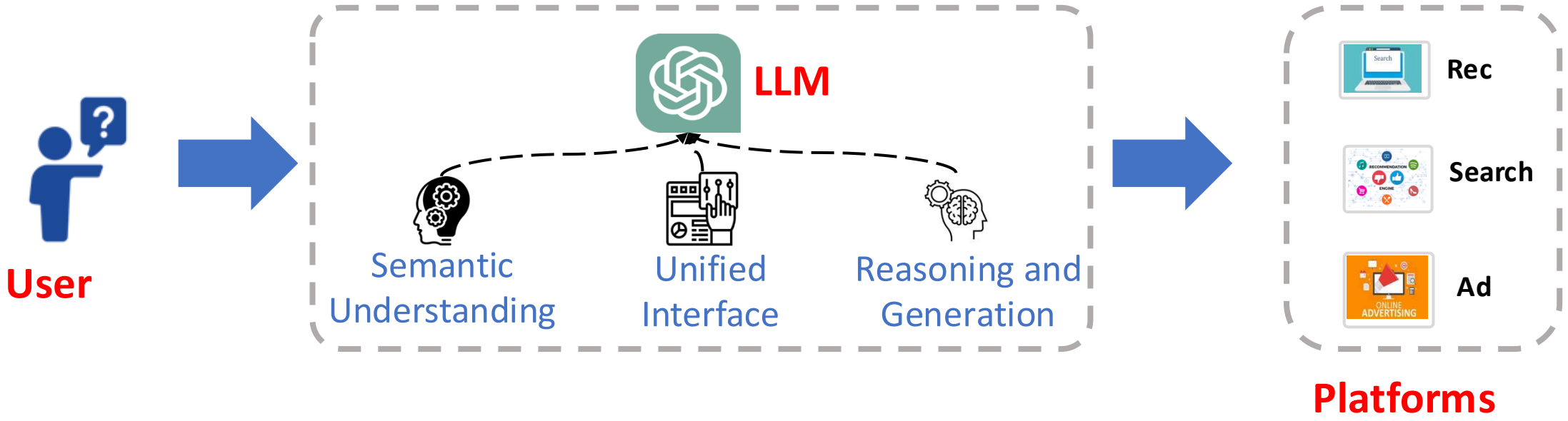
Narrow Scope



- Only serves a single scenario and lack cross-scenario collaboration
- Lack proactive service to users

LLM Driven Personalization

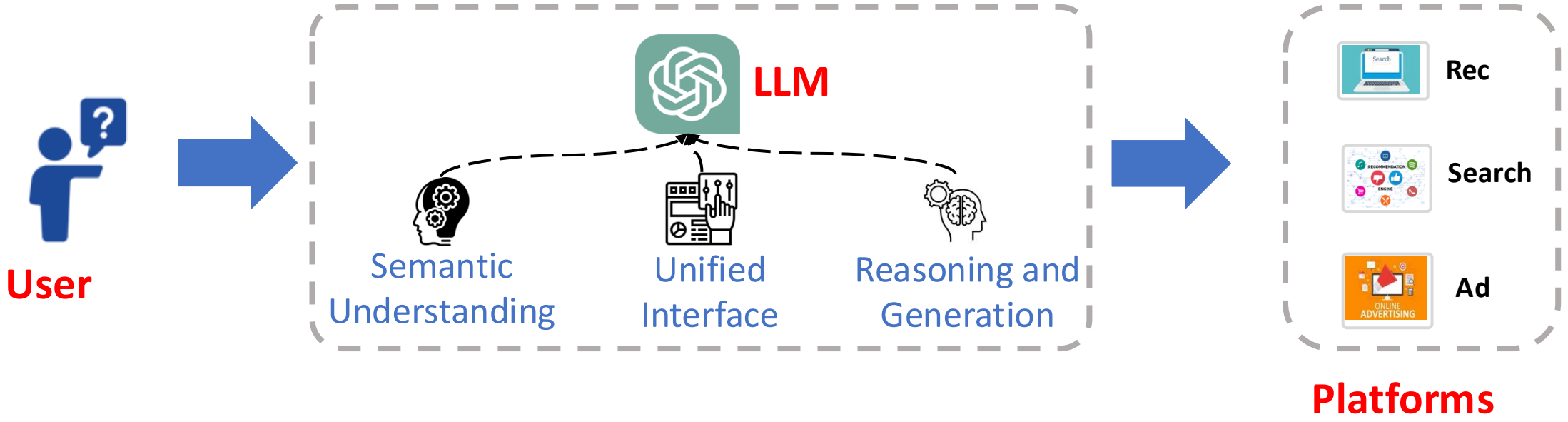
➢LLM Driven Personalization Pipeline



LLM Driven Personalization Limitations

- Outdated Knowledge & Hallucination
- Passive & Stateless
- No Action & Tool Integration

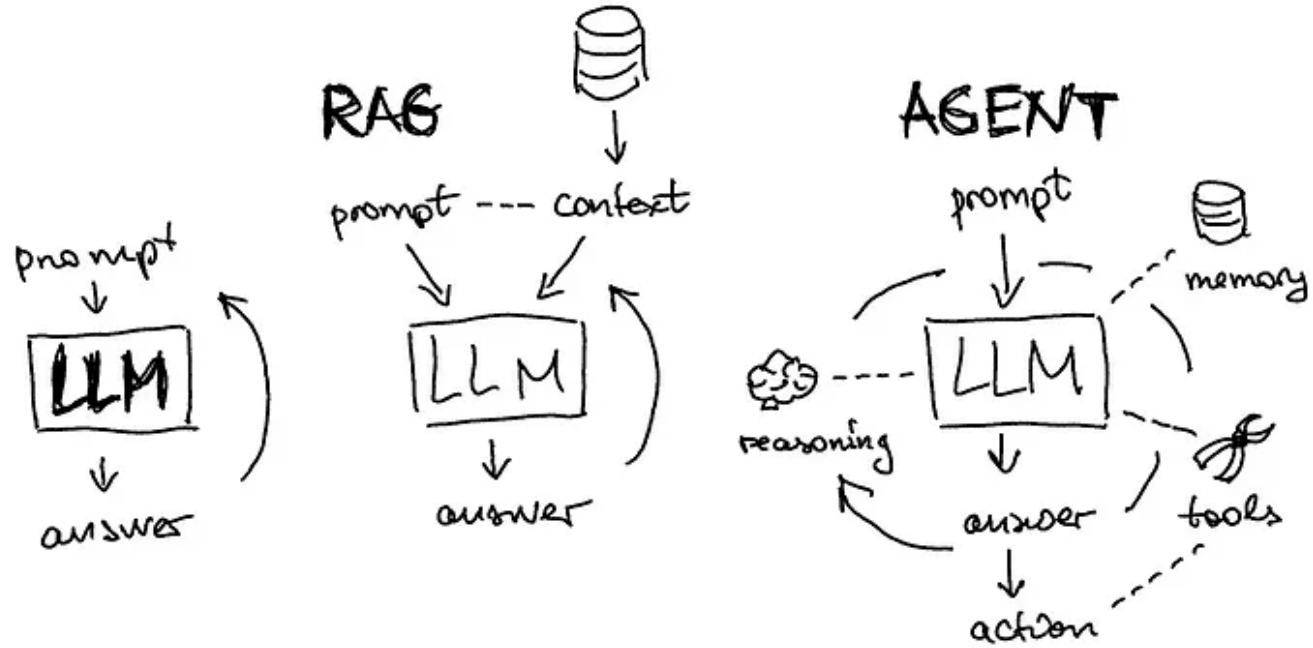
Answer is RAG and Agent!



Introduction

Questions:

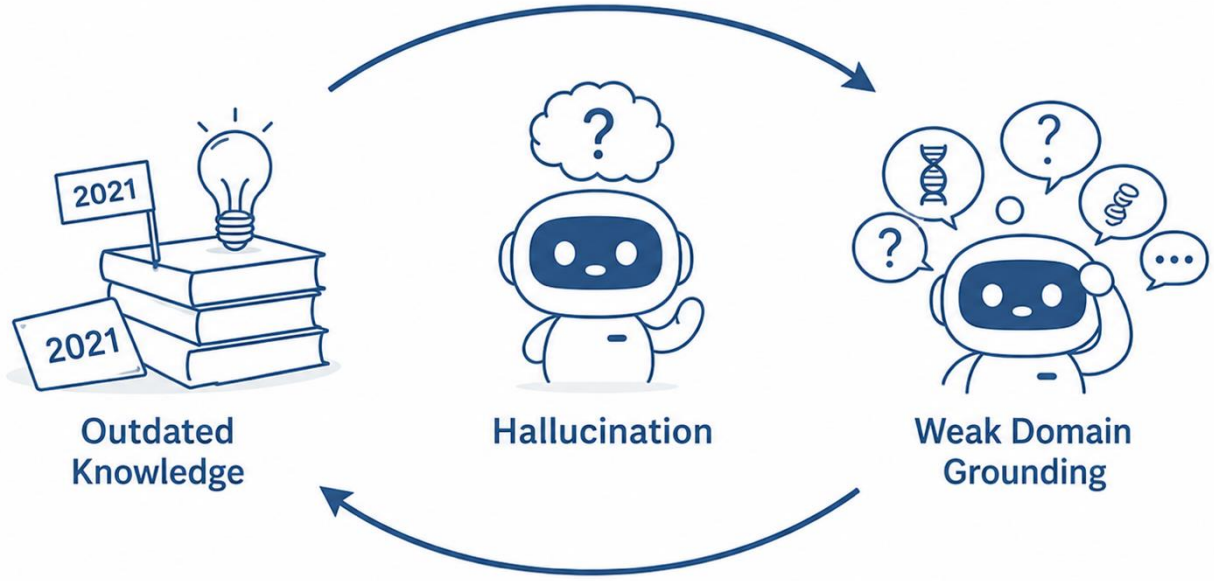
- What is RAG?
- What is Agent / Agentic RAG?
- What is Personalization?
- What is Personalization in RAG and Agent?



Introduction – Why RAG

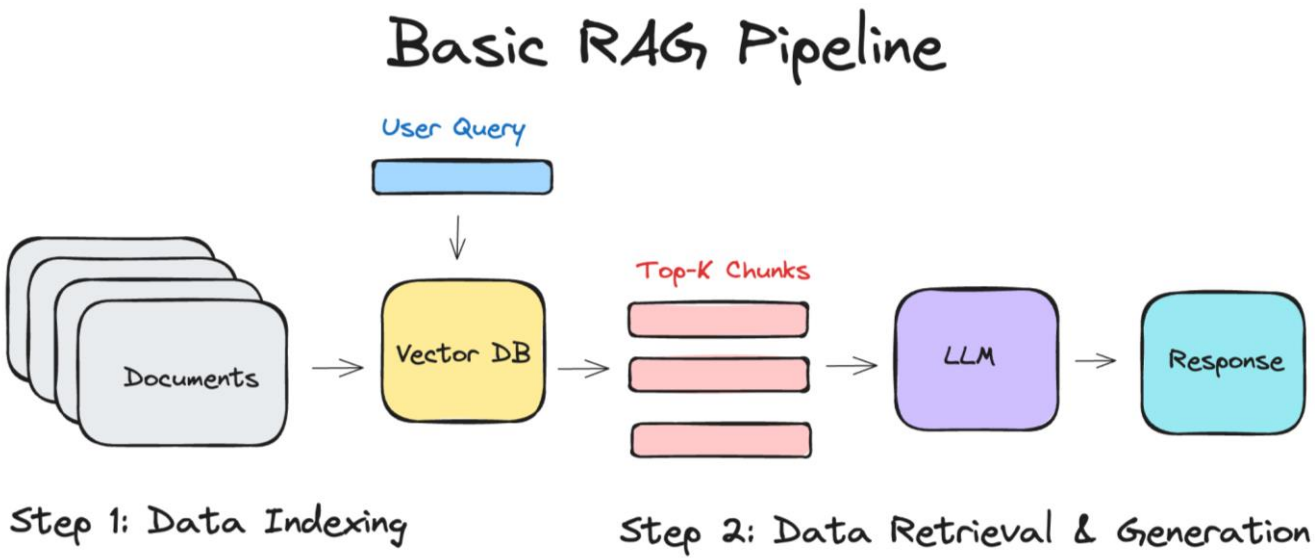
- LLMs are powerful but **not always grounded**

- **Key issues:**
  - **Outdated knowledge**
  - **Hallucination**
  - **Weak domain grounding**



Introduction – What is RAG

- RAG: **retrieve external evidence before generation**
- A RAG system can be decomposed into pre-retrieval, retrieval, and generation.
- Pre-retrieval
  - Understand, rewrite, expand, route
- Retrieval / post-retrieval
  - Index, search, rank, compress
- Generation
  - Synthesize, cite, format, answer



Introduction – Beyond RAG

What RAG Adds and What It Still Misses?

- RAG helps:
  - **Fresh knowledge**
  - **Factual grounding**
  - **Domain adaptation**

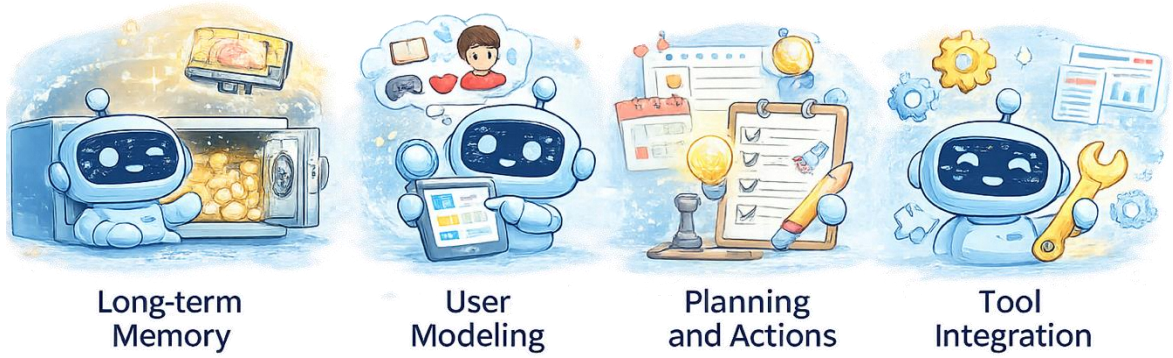


Introduction – Beyond RAG

What RAG Adds and What It Still Misses?

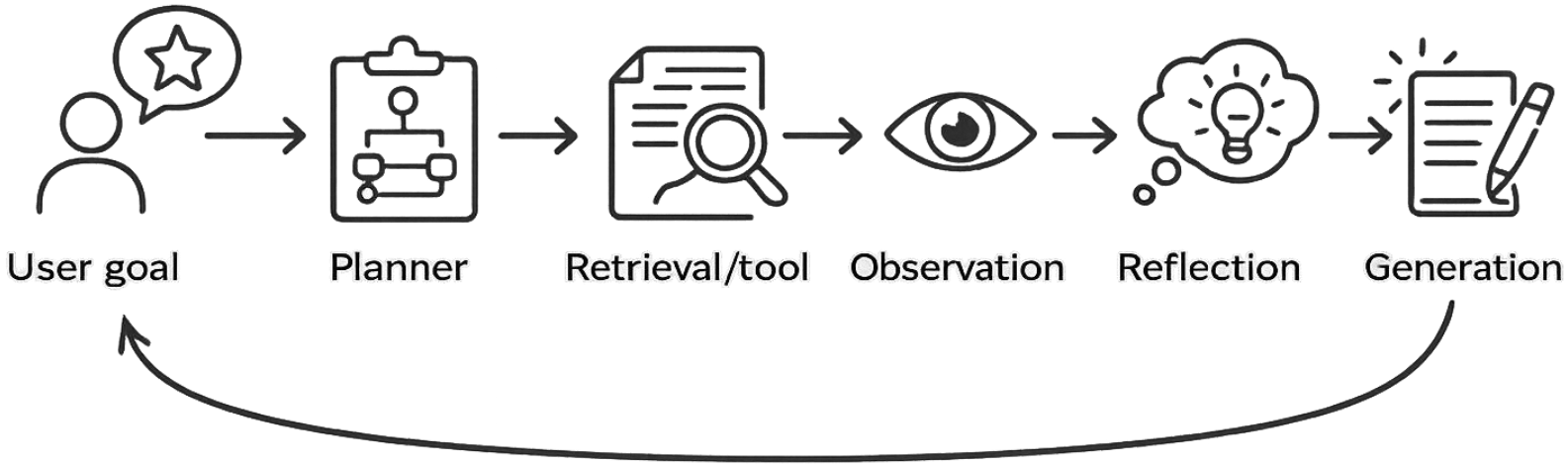
- RAG helps:
  - **Fresh knowledge**
  - **Factual grounding**
  - **Domain adaptation**
- Still missing:
  - **Long-term memory**
  - **User modeling**
  - **Planning and actions**
  - **Tool integration**





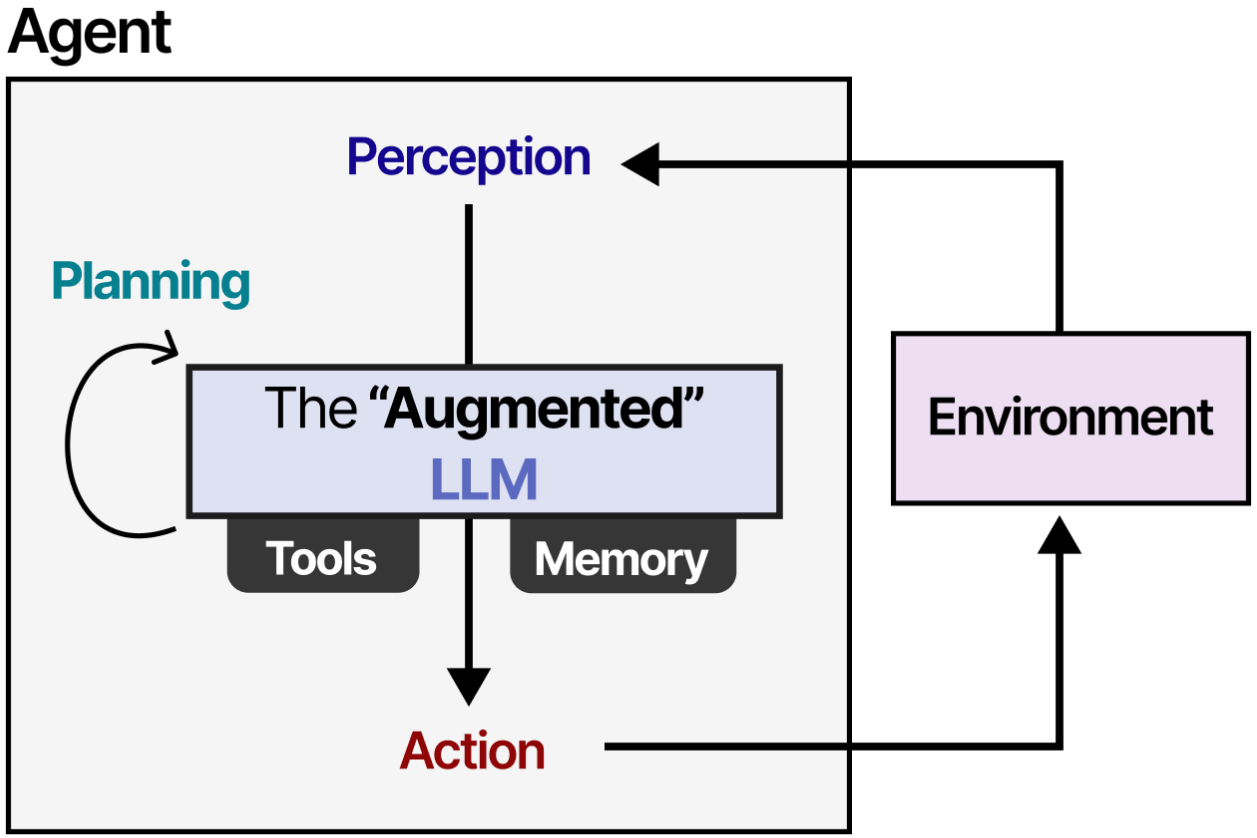
From RAG to Agentic RAG

- Standard RAG: retrieve -> generate
- Agentic RAG: plan -> retrieve/tool -> observe -> revise -> answer
- More control over:
  - When to retrieve
  - What to retrieve
  - Which tools to use
  - When to stop



What Makes an LLM System an Agent?

- Goal-directed behavior (**Planning**)
- Persistent memory (**Memory**)
- Planning and tool/API use (**Tool Use**)



RAG and Agent Workflows Align

RAG

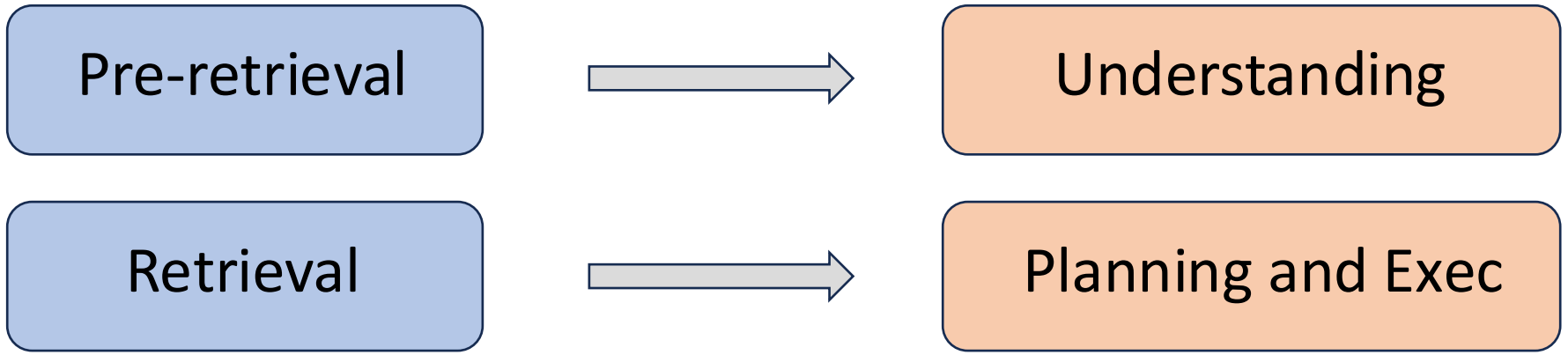
Pre-retrieval



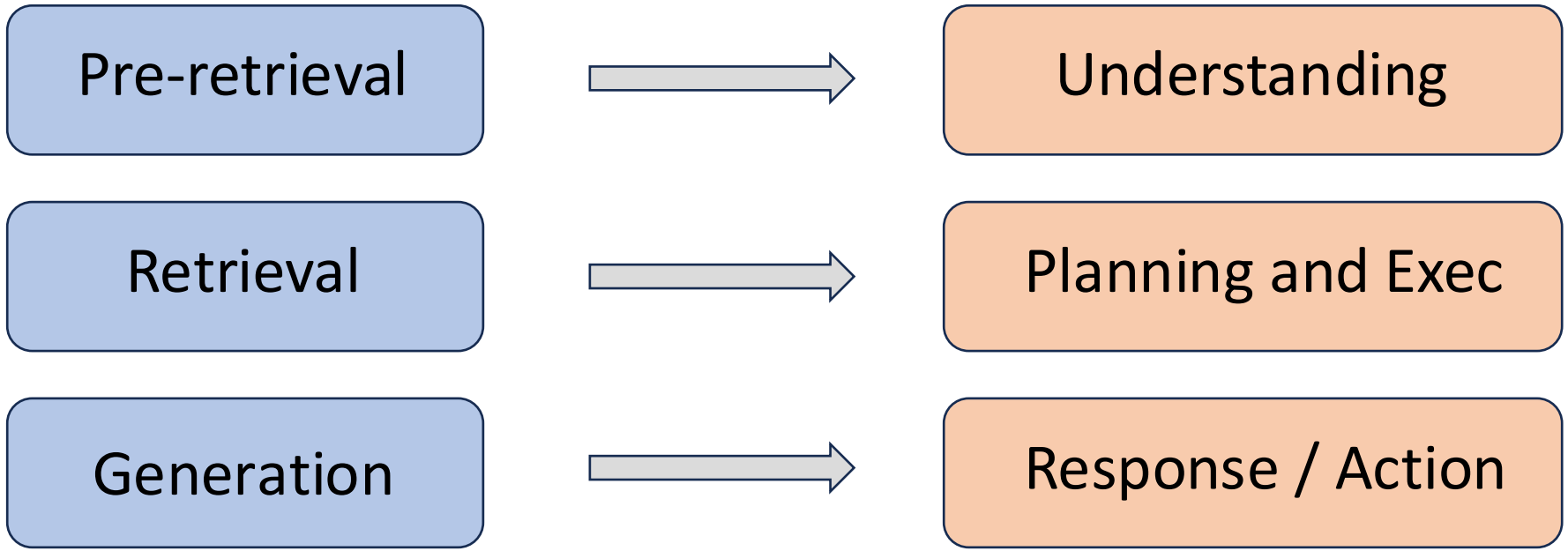
Agent

Understanding

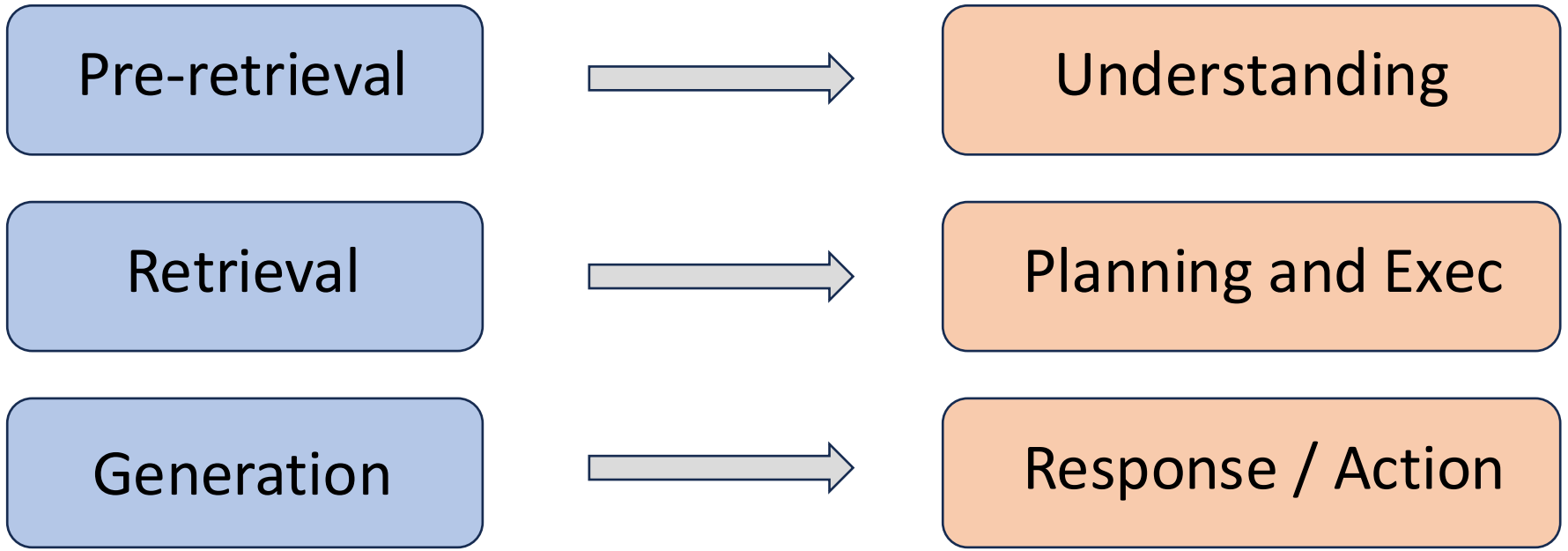
RAG and Agent Workflows Align



RAG and Agent Workflows Align



RAG and Agent Workflows Align

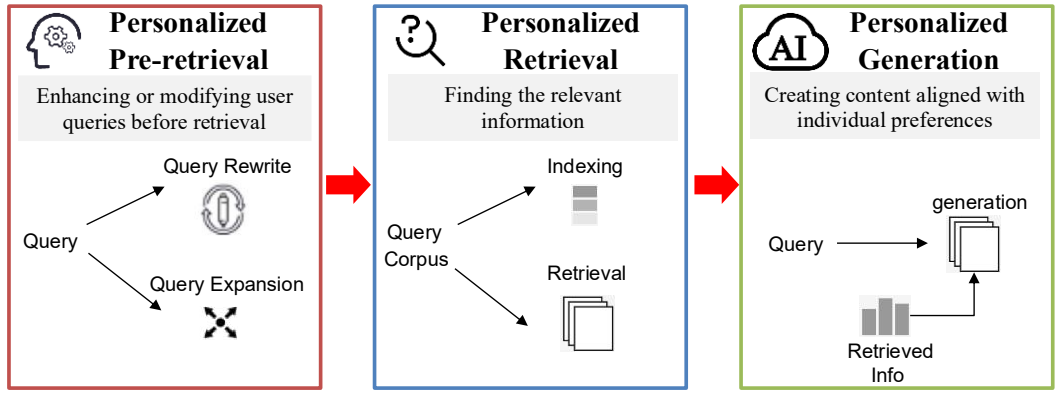


RAG is a bridge to Agents!

How to use RAG / Agent for Personalization?

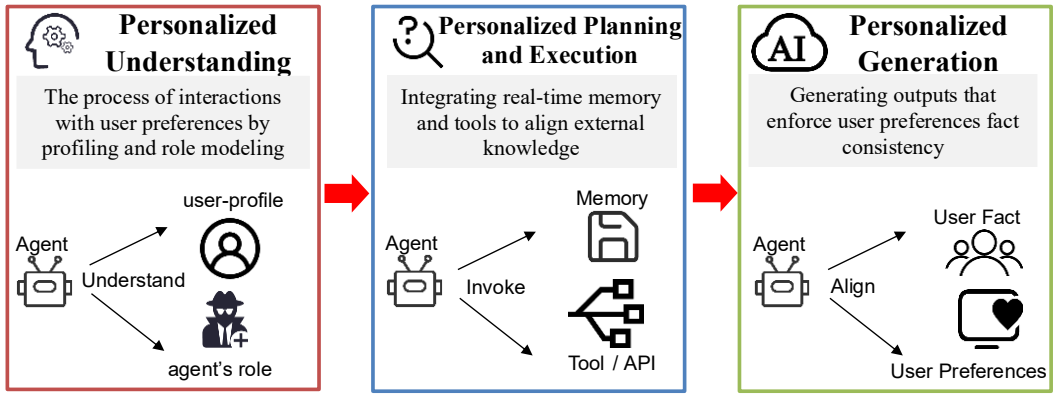
RAG/Agent Driven Personalization

RAG Flow



LLM + Personalization

Agent Work Flow



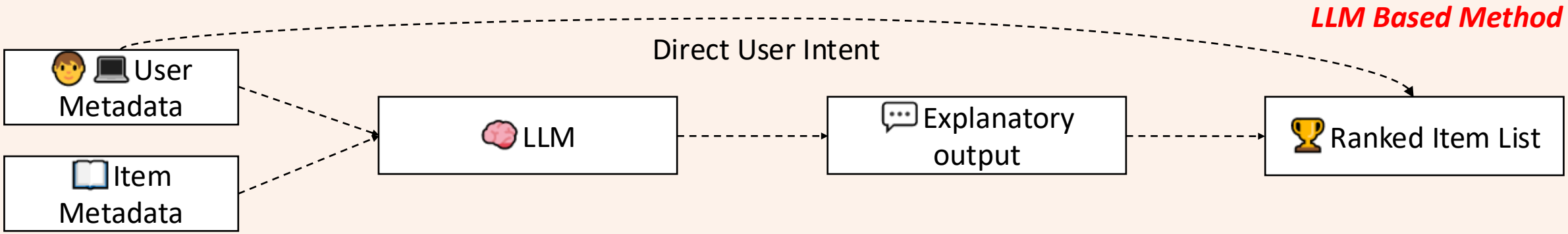
- RAG/Agent + Personalization
  - Next-generation personalized system

- RAG + Personalization
  - **Up-to-date** information
  - Boosts **factual accuracy**
  - Reduces **hallucinations**

- Agent + Personalization
  - Enhanced Capabilities including **Memory, Planning and Tool Usage**
  - Enabling **proactive, action-oriented** service

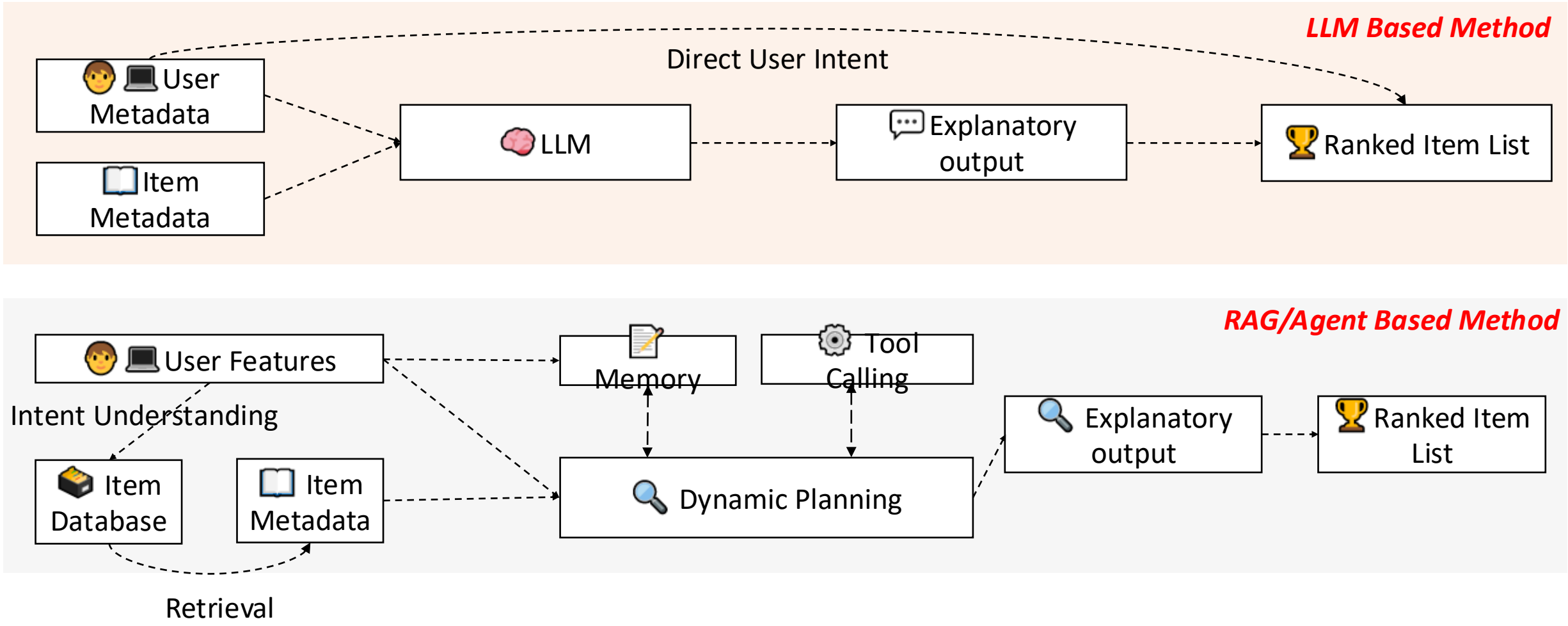
RAG/Agent Driven Personalization

➢Case: Recommend Shopping Items for Users



RAG/Agent Driven Personalization

➢Case: Recommend Shopping Items for Users

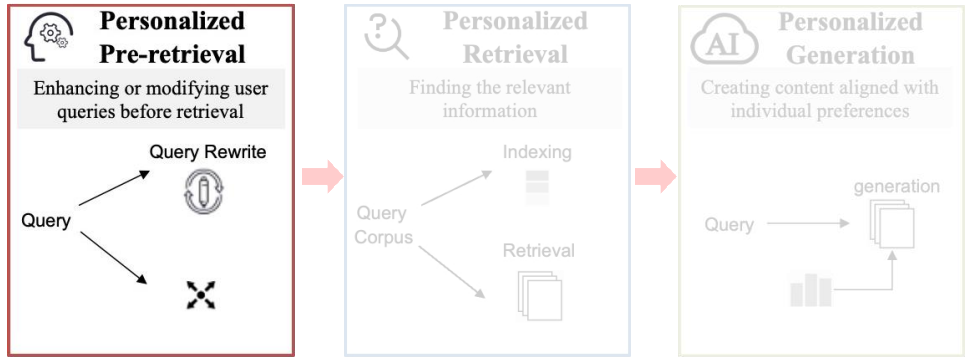


How to adopt Personalization in RAG/Agent

➢Pre-retrieval

➢**Query Processing Q** refines the query q using personalized information p, such as through query rewriting or expansion

RAG Flow



Agent Work Flow

$$g = \mathcal{G}\left(\mathcal{R}\left(\underbrace{\mathcal{Q}\left(q, p\right)}_{\text{Pre-retrieval}}, \mathcal{C}, p\right), \text{prompt}, p, \theta\right)$$

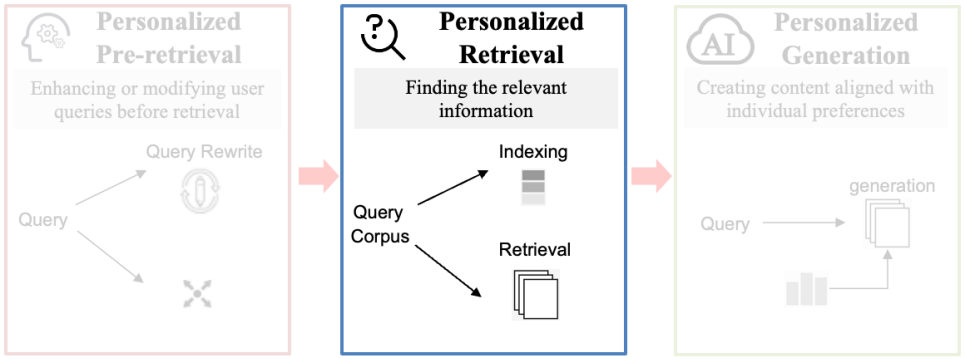
The Personalized Info p

- **Explicit user profile:**
  - attributes and stated preferences, Age, Occupation, Preferences
- **Historical interactions**
  - clicks, searches, purchases, dialogue turns
- **Historical content**
  - reviews, notes, documents, chat history
- **Persona / role signals**
  - expertise, role, simulated persona
- **Key tradeoff: interpretability** vs. **sparsity, noise, and privacy**

How to adopt Personalization in RAG/Agent

- **Retrieval**
  - **Retriever R** leverages personalization p to fetch relevant documents from the corpus C

RAG Flow



Agent Work Flow

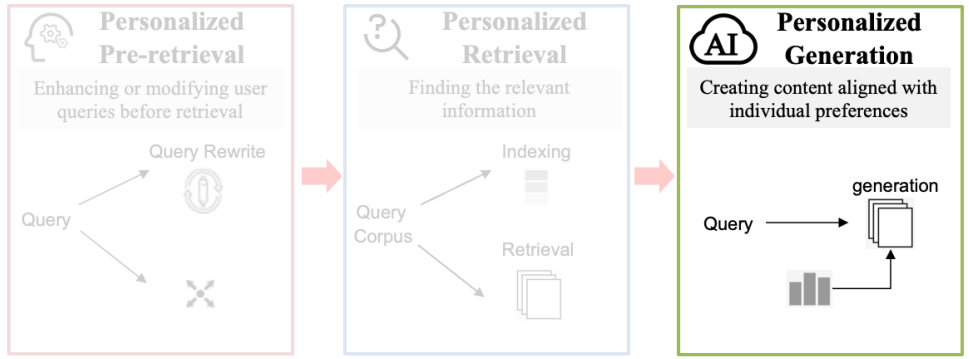
$$g = \mathcal{G}\left(\mathcal{R}\left(\mathcal{Q}\left(q, p\right), C, p\right), \text{prompt}, p, \theta\right)$$

Retrieval

How to adopt Personalization in RAG/Agent

- **Generation**
  - The retrieved information, combined with personalization p and structured using the given prompt, fed into the generator G whose parameter is θ.

RAG Flow



Agent Work Flow

$$g = \underbrace{\mathcal{G}\left(\mathcal{R}\left(\mathcal{Q}\left(q, p\right), \mathcal{C}, p\right), \text{prompt}, p, \theta\right)}_{\text{Generation}}$$

How to adopt Personalization in RAG/Agent

- **Agent Work Flow**
  - We consider the **agent system as a specialized application of the RAG framework**, where personalization is incorporated in a manner similar to the RAG framework.

RAG Flow

Agent Work Flow

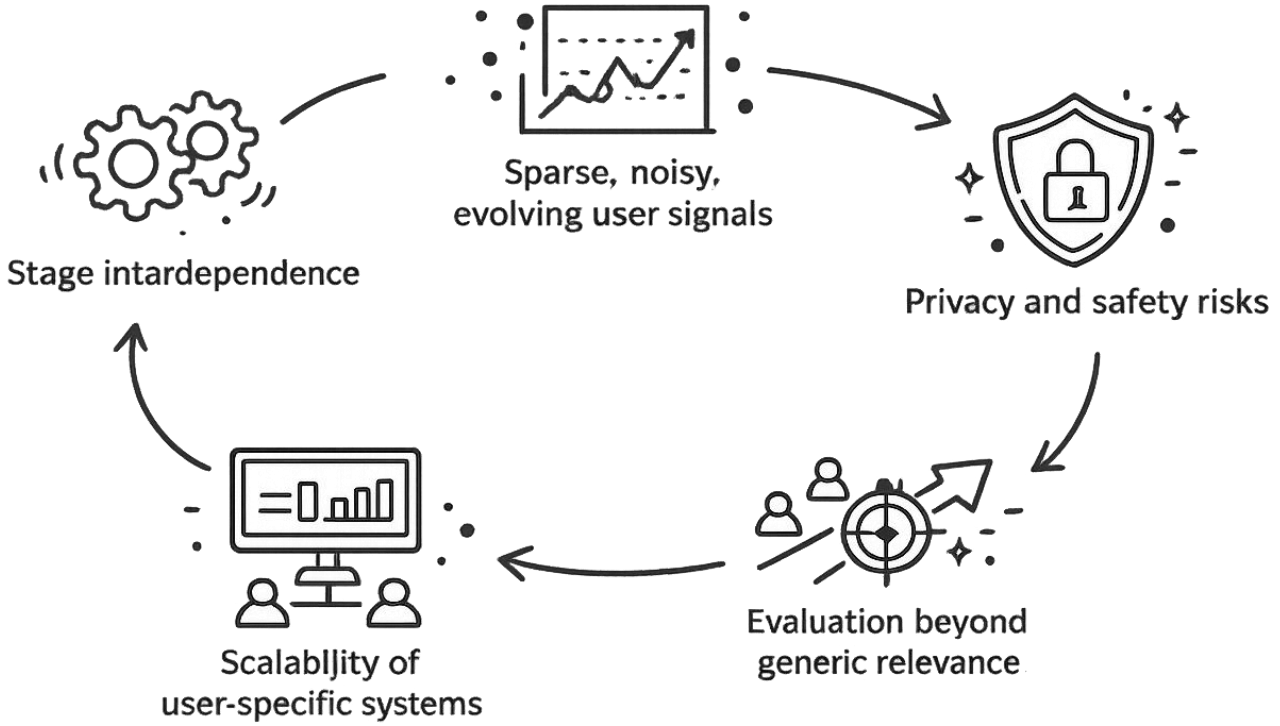


$$g = \mathcal{G}\left(\mathcal{R}\left(\mathcal{Q}\left(q, p\right), \mathcal{C}, p\right), \text{prompt}, p, \theta\right)$$

Agent Work Flow

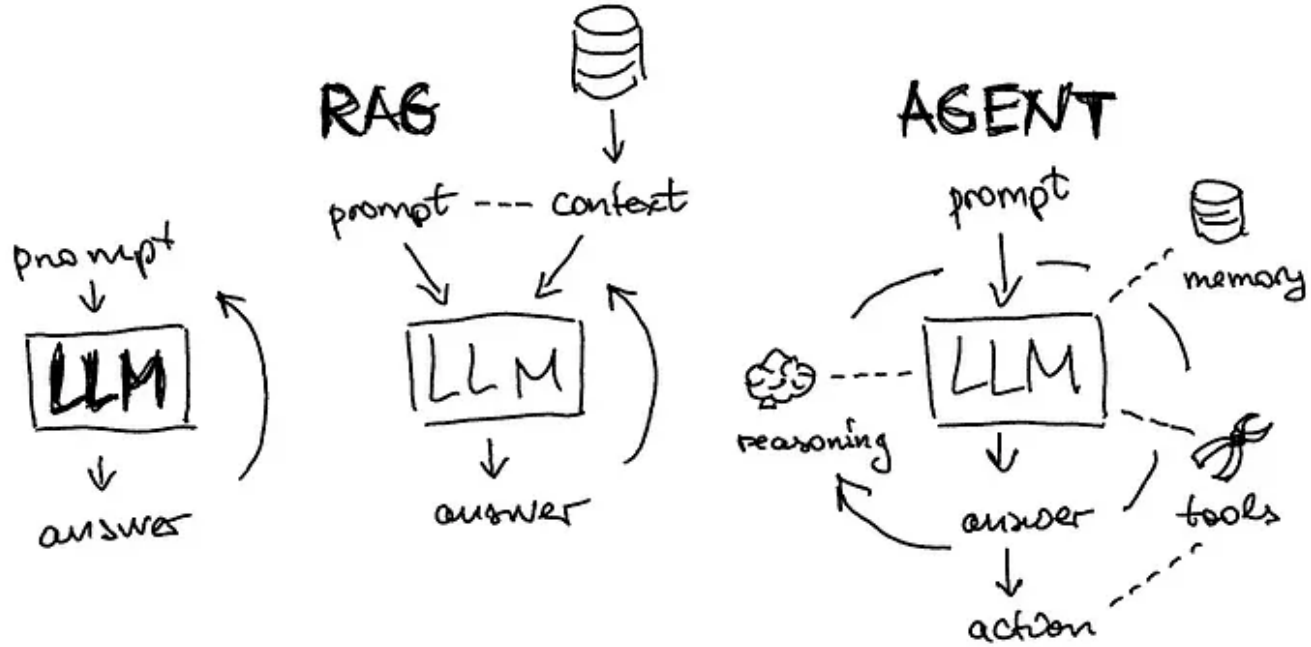
What Makes Personalization Hard?

- Stage **interdependence**
- **Sparse, noisy, evolving** user signals
- **Privacy** and **safety** risks
- Evaluation **beyond generic relevance**
- **Scalability** of user-specific systems



Organization of This Tutorial

- **Personalized RAG**
  - Pre-retrieval
  - Retrieval / post-retrieval
  - Generation
- **Personalized agents**
  - Understanding
  - Planning and execution
  - Generation
- **Evaluation, datasets, challenges, future directions**



Agenda

1. Introduction
9:00 – 9:50



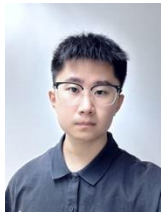
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

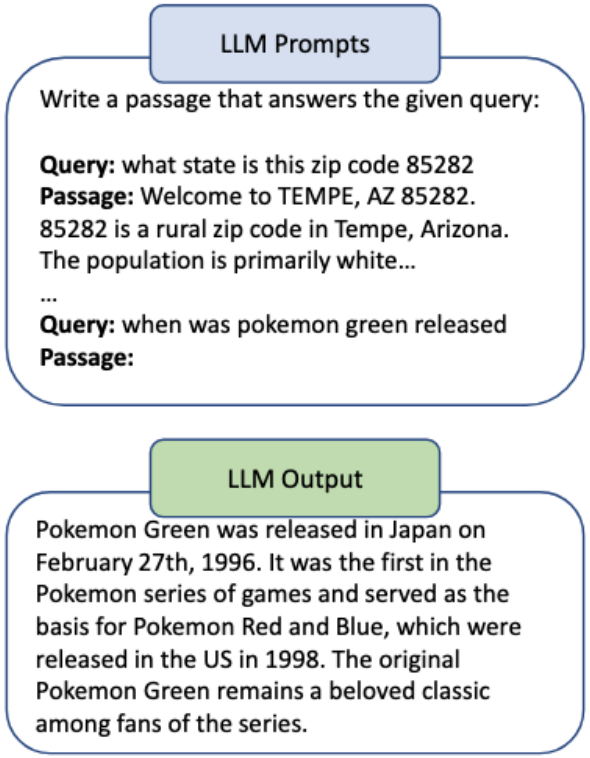
8. Hands on Projects
16:45 – 17:30



Pengyue Jia

Pre-retrieval: Problem

Existing pre-retrieval methods rely on LLM world knowledge to complete query semantics for retrieval over general corpora



Query2Doc(EMNLP, Google)

Pre-retrieval: Problem

Existing pre-retrieval methods rely on LLM world knowledge to complete query semantics for retrieval over general corpora

LLM Prompts

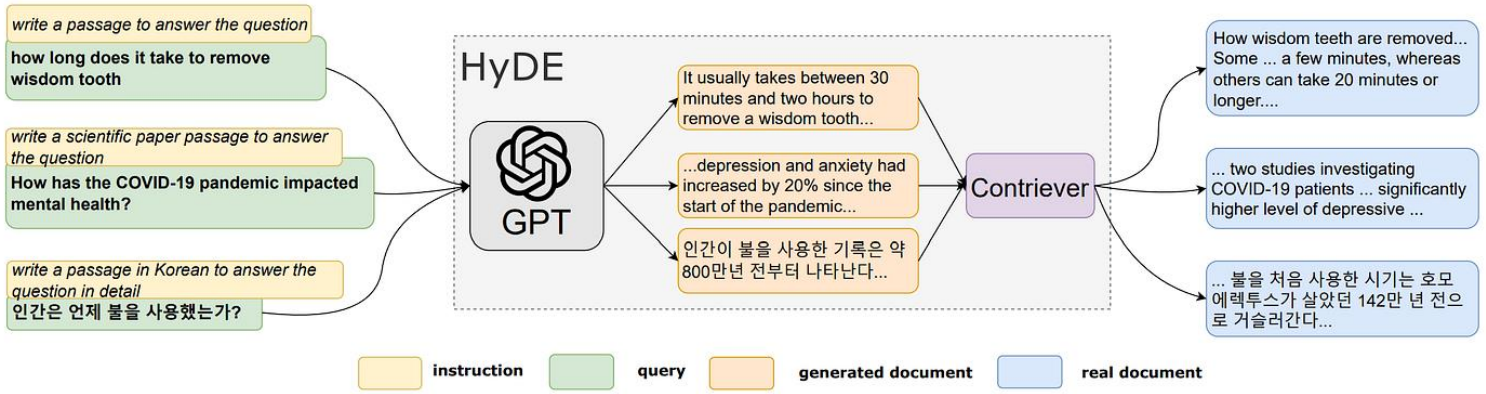
Write a passage that answers the given query:

Query: what state is this zip code 85282
Passage: Welcome to TEMPE, AZ 85282. 85282 is a rural zip code in Tempe, Arizona. The population is primarily white...
...
Query: when was pokemon green released
Passage:

LLM Output

Pokemon Green was released in Japan on February 27th, 1996. It was the first in the Pokemon series of games and served as the basis for Pokemon Red and Blue, which were released in the US in 1998. The original Pokemon Green remains a beloved classic among fans of the series.

Query2Doc(EMNLP, Google)



HyDE(ACL, CMU)

Pre-retrieval: Problem

Existing pre-retrieval methods rely on LLM world knowledge to complete query semantics for retrieval over general corpora

LLM Prompts

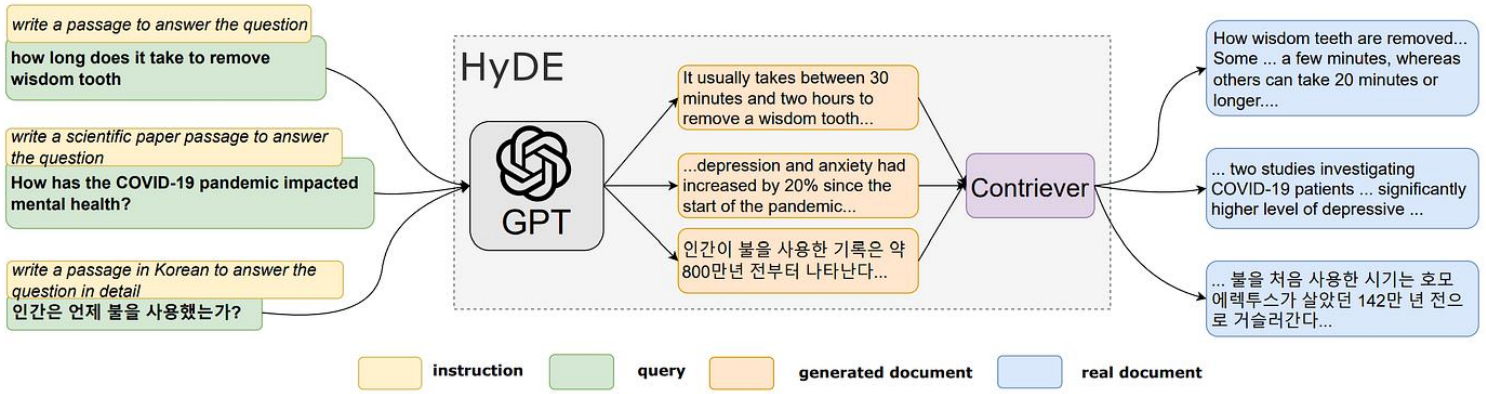
Write a passage that answers the given query:

Query: what state is this zip code 85282
Passage: Welcome to TEMPE, AZ 85282. 85282 is a rural zip code in Tempe, Arizona. The population is primarily white...
...
Query: when was pokemon green released
Passage:

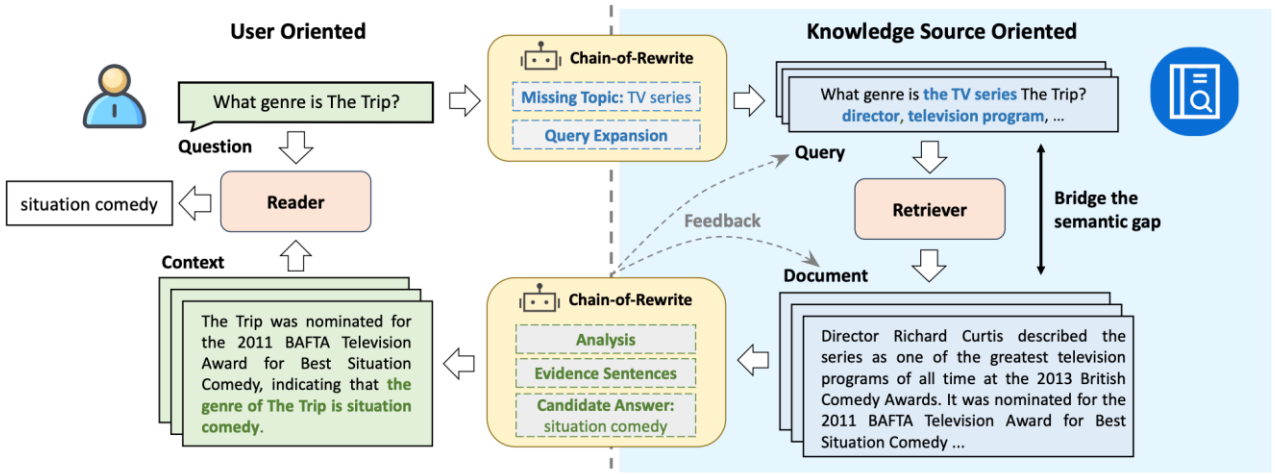
LLM Output

Pokemon Green was released in Japan on February 27th, 1996. It was the first in the Pokemon series of games and served as the basis for Pokemon Red and Blue, which were released in the US in 1998. The original Pokemon Green remains a beloved classic among fans of the series.

Query2Doc(EMNLP, Google)



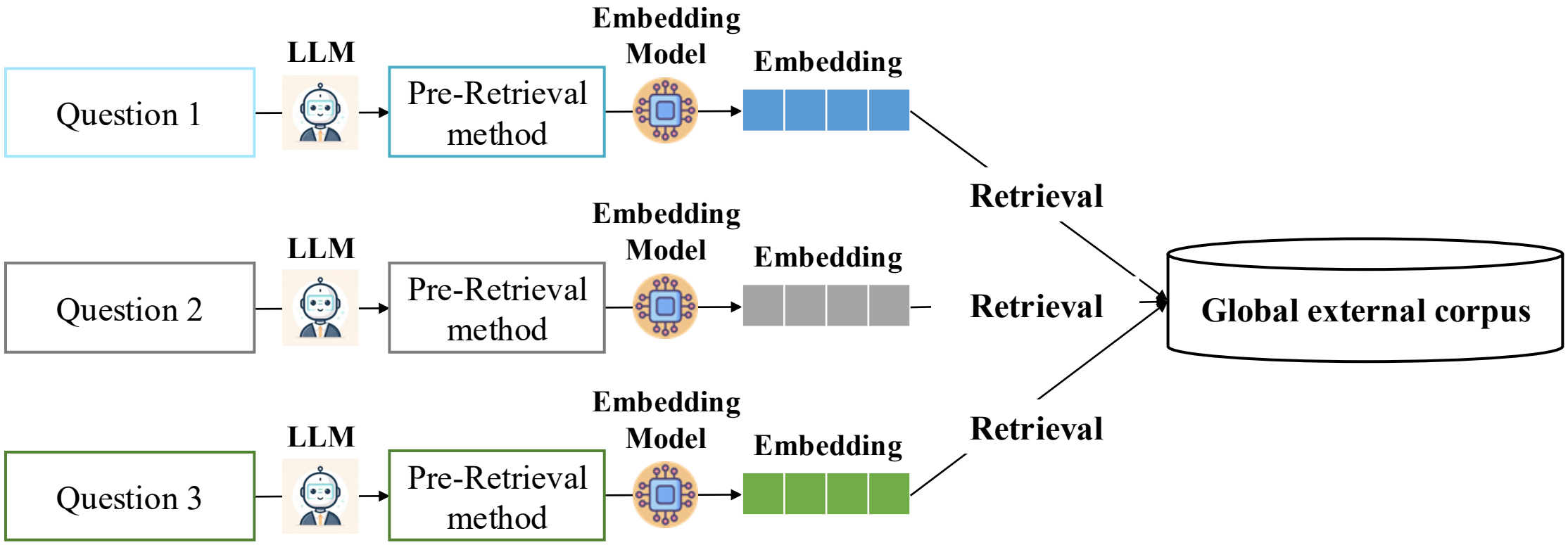
HyDE(ACL, CMU)



Chain-of-Rewrite (EMNLP, UCAS and Ant)

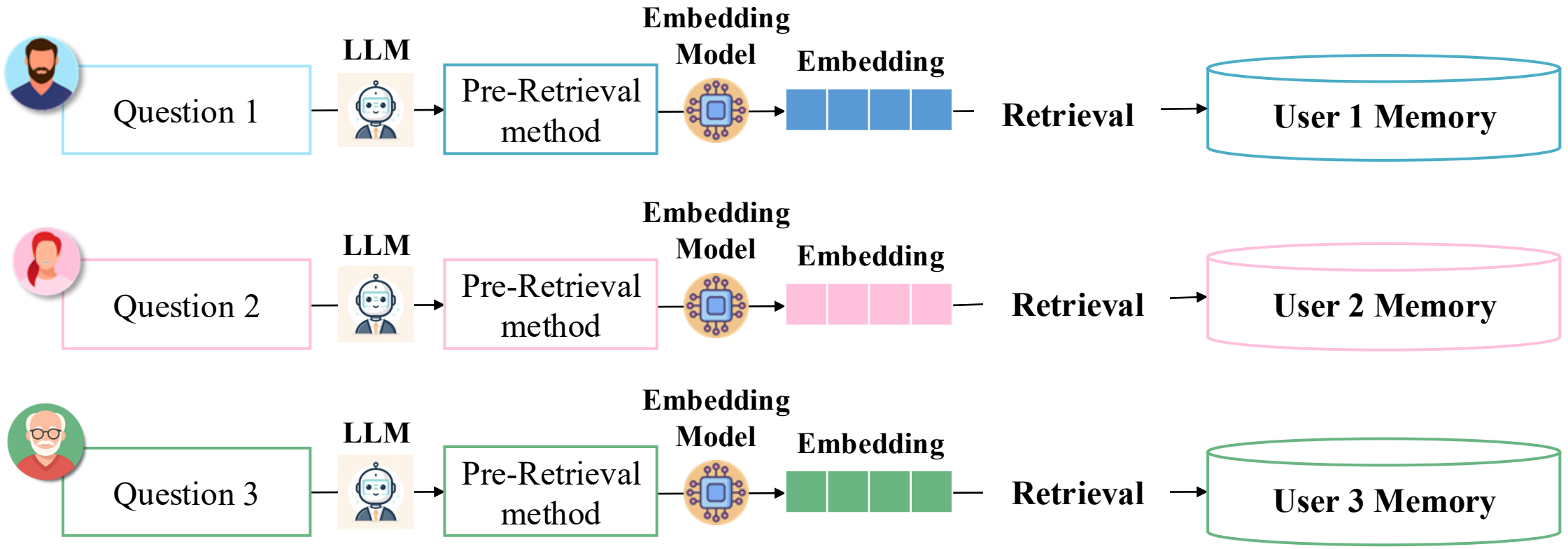
Pre-retrieval: Problem

Assuming a shared external corpus



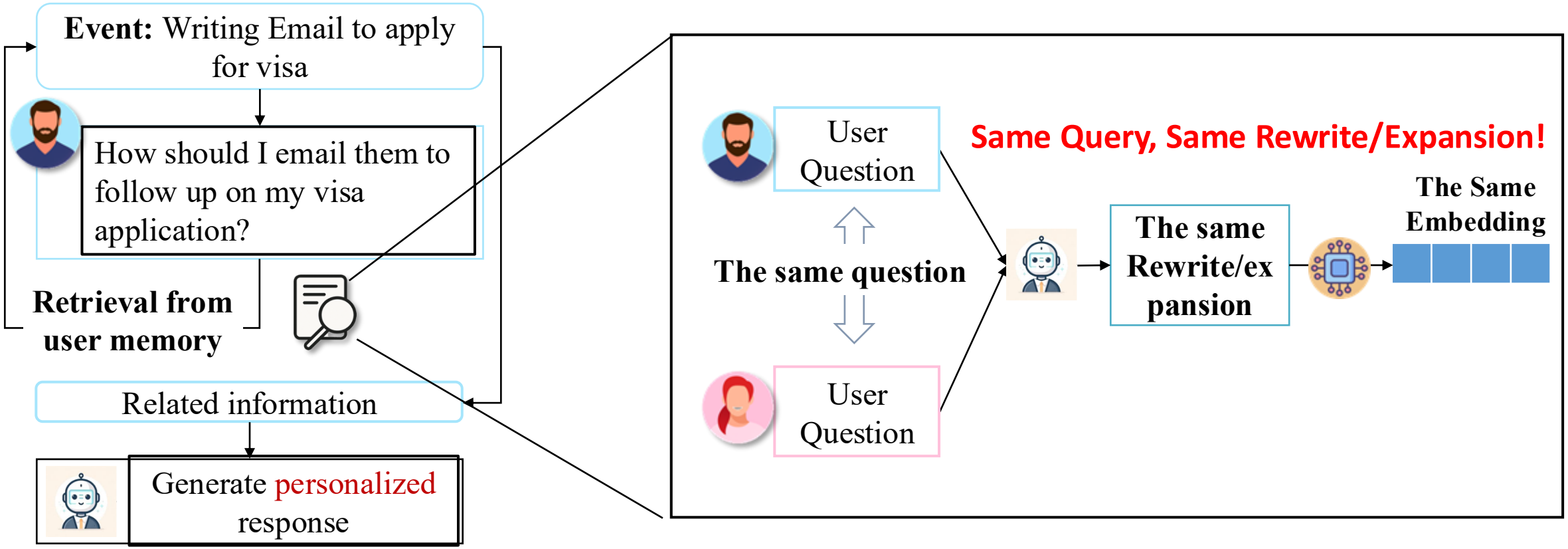
Pre-retrieval: Problem

However, in personalized scenario, it is required to retrieval over **user personal copra**



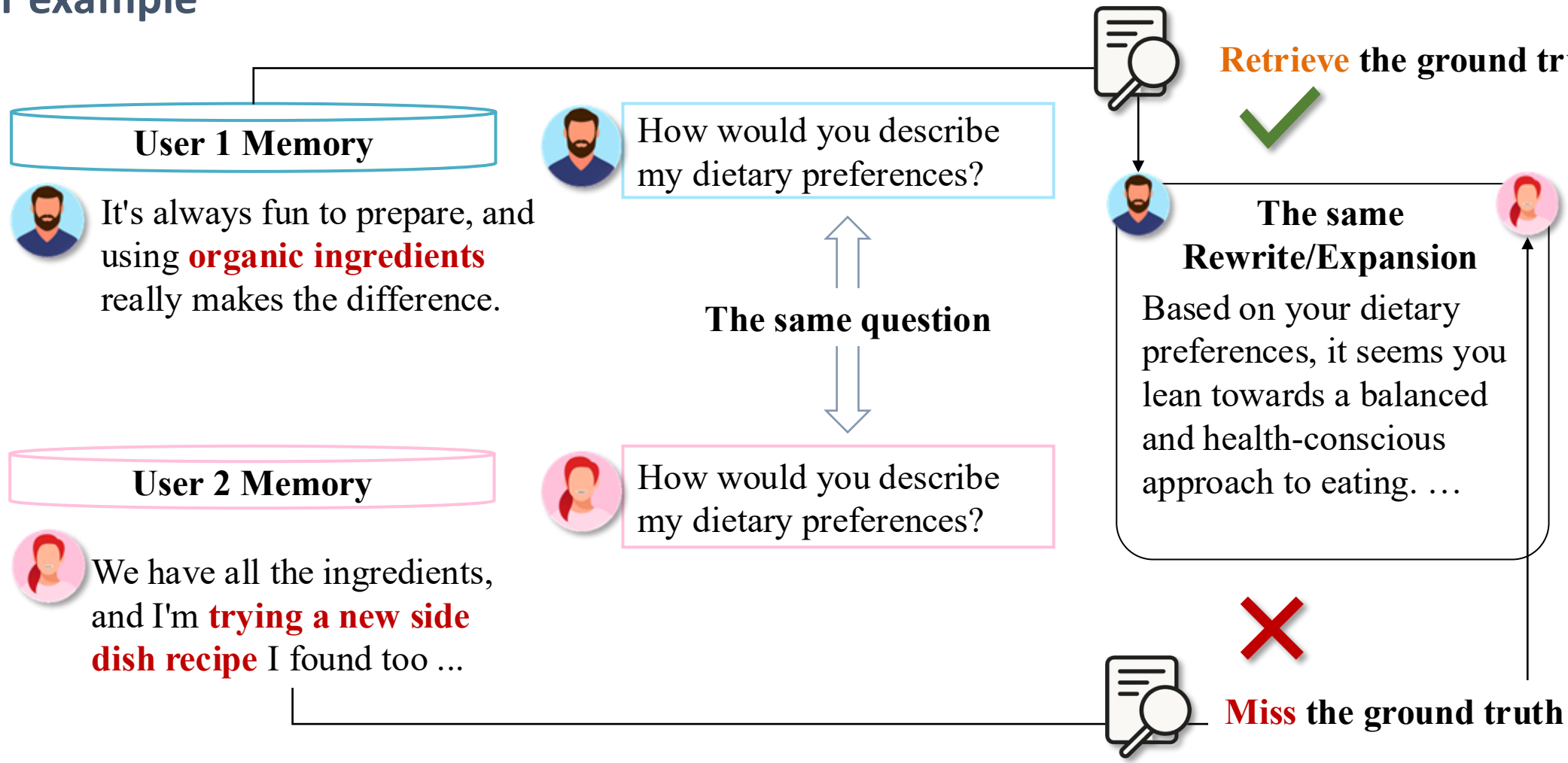
Pre-retrieval: Problem

When apply these method in personalized scenario, it would be:



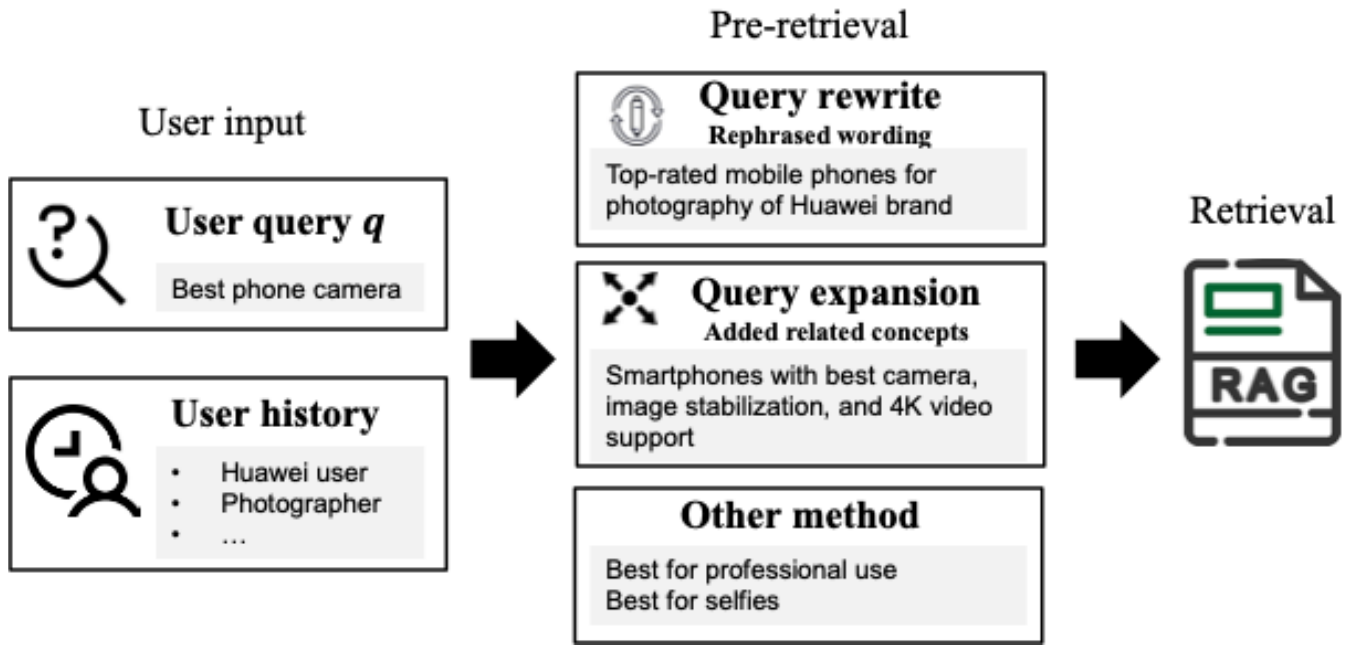
Pre-retrieval: Problem

For example



Pre-retrieval: Formula definition

- Pre-retrieval Definition:
  - ***The original user query is enhanced or modified by using personalized information*** before the retrieval process to improve the relevance and quality of the search results.

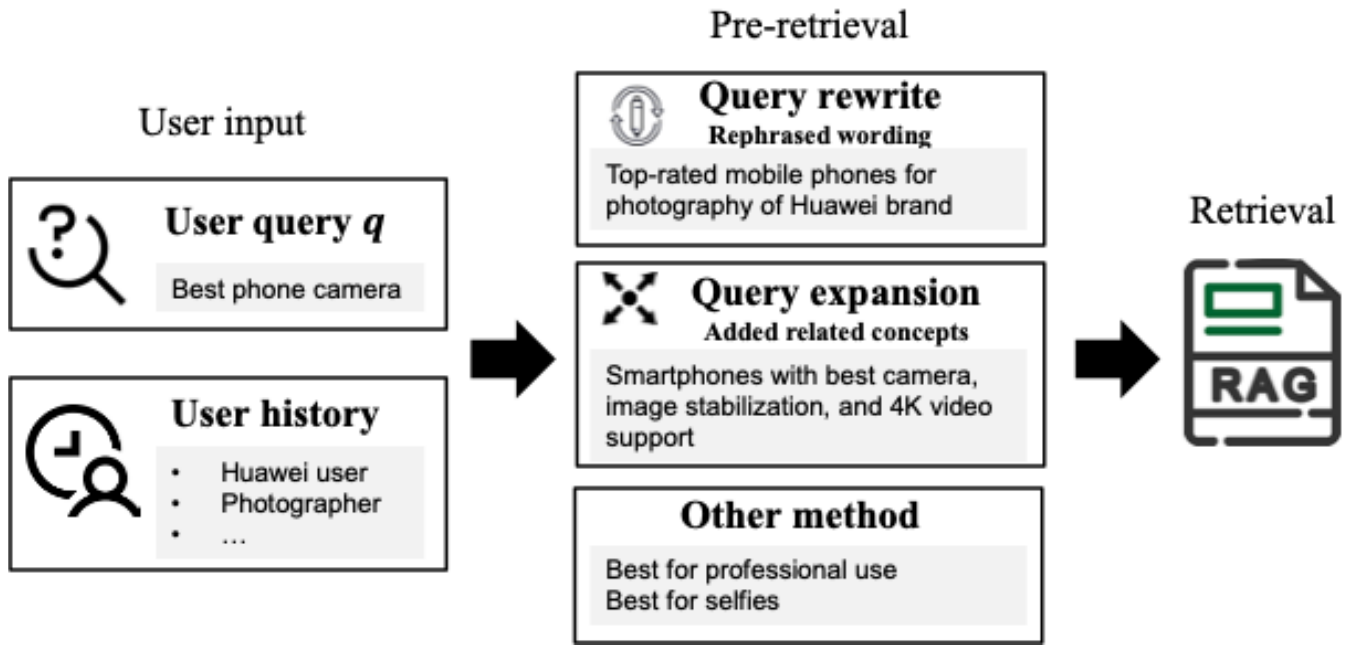


Pre-retrieval: Formula definition

- Pre-retrieval Definition:
  - ***The original user query is enhanced or modified by using personalized information*** before the retrieval process to improve the relevance and quality of the search results.

- Equation: $q^* = Q(q, p)$
  - q is the user query
  - p is the personalization information of the user
  - Q is the query process
  - q^* is the enhanced query

- According to Q, methods catrgories:
  - Query Rewrite
  - Query Expansion
  - Else



Pre-retrieval: Formula definition

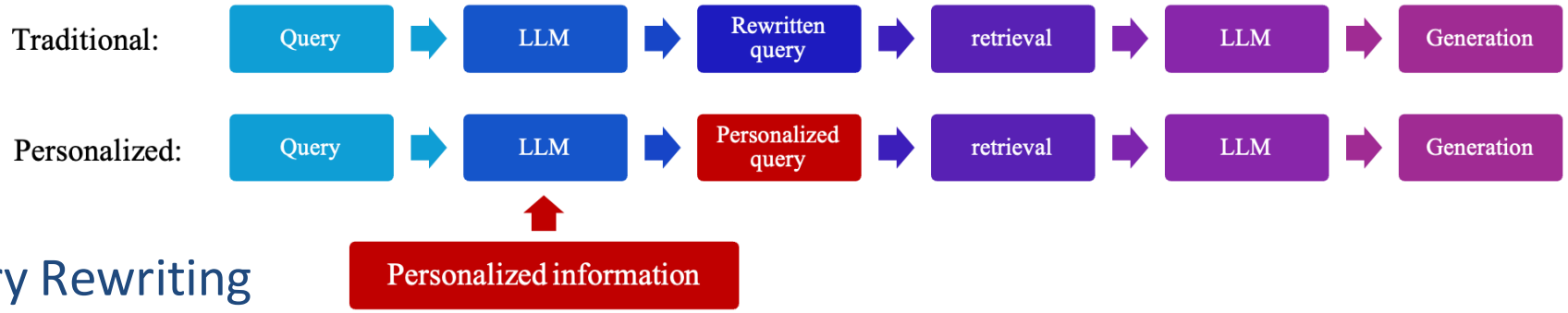
Method	Description	Pros	Cons
		Query Rewriting	
Direct Query Rewriting	Reformulates using personalized methods	• Simple implementation • Handles semantic errors	• Needs extensive user data • Limited domain adaptation • Risk of overfitting
Auxiliary Query Rewriting	Uses external tools: retrieval, reasoning, memory	• Uses multiple sources • Better complex reasoning • More flexible and robust	• High computational cost • Increased latency
		Query Expansion	
Tagging-based Expansion	Expands using social tags and metadata	• Collective intelligence • Dynamic trend adaptation • Rich semantic relations	• Depends on tag quality • Vulnerable to noise
History/Network-based Query Expansion	Uses user history, networks, embeddings or profiles	• Captures semantic similarity • Language-agnostic	• Compute intensive • Potential semantic drift • Hard to interpret
		Others	
Other Methods	Query disambiguation and auto-completion	• Improves user experience • Real-time guidance • Prevents early errors	• Limited scope • Privacy concerns • Constrains expression

Pre-retrieval: Query Rewrite

- Motivation
  - ***Traditional rewrites fail to capture user-specific intent and semantic variation.***

- Personalized Query Rewrite Definition
  - Query rewriting in RAG at the pre-retrieval stage refers to ***the process of reformulating user queries to enhance retrieval effectiveness by improving relevance, disambiguating intent, from personalized information*** before retrieving documents from an external knowledge source.

- Methods can be divided to:
  - Direct Personalized Query Rewriting
  - Auxiliary Personalized Query Rewriting

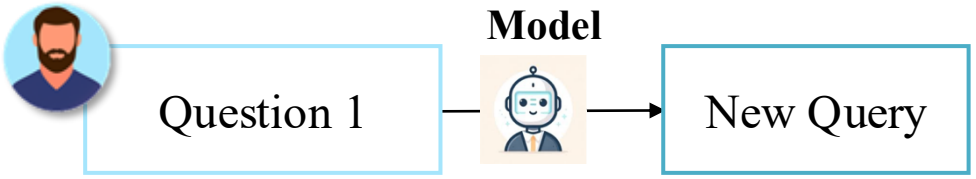


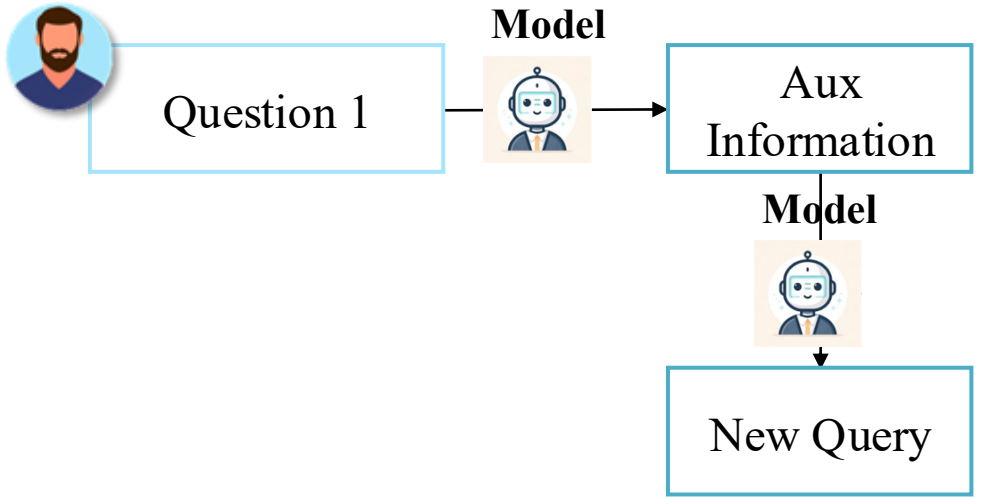
Pre-retrieval: Query Rewrite

Query Rewrite

Direct Personalized Query Rewriting

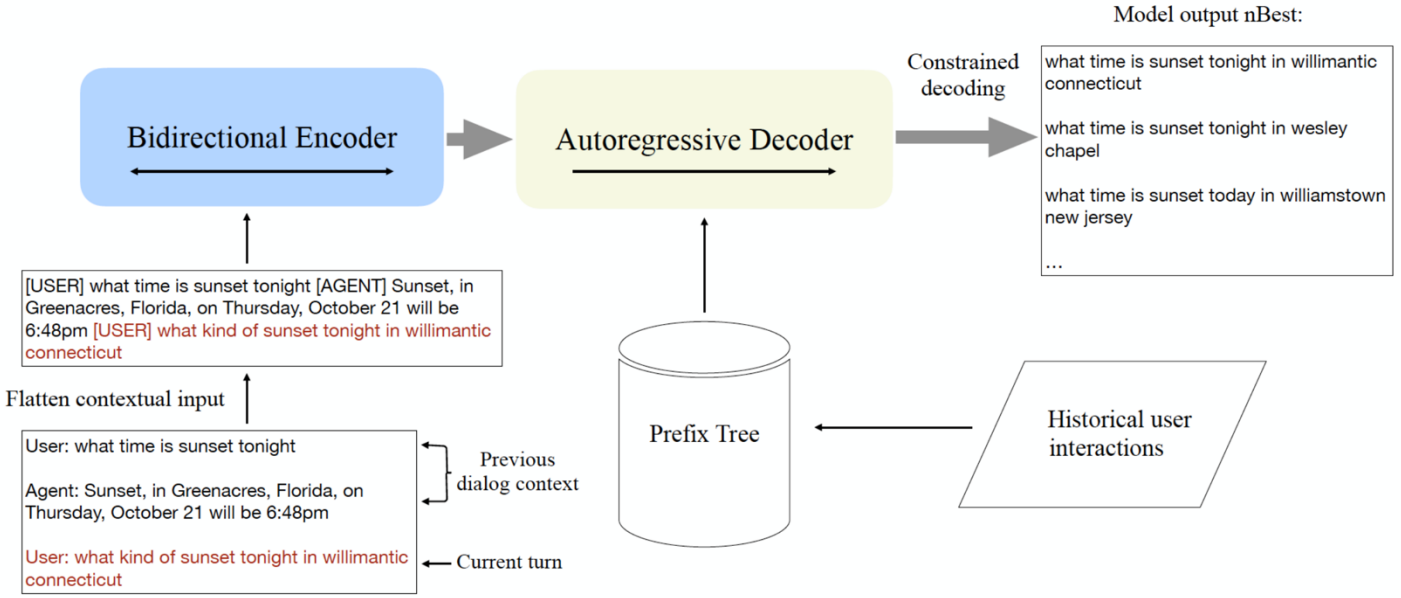
Auxiliary Personalized Query Rewriting





Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models
- CGF(EMNLP'22)
  - **Problem:**
    - User query are incomplete or have noise
    - Unconstrained generation often leads to unstable, or hallucinated rewrites
  - **QR method Require**
    - preserves user intent,
    - incorporates relevant context,
    - and is suitable for downstream retrieval systems.

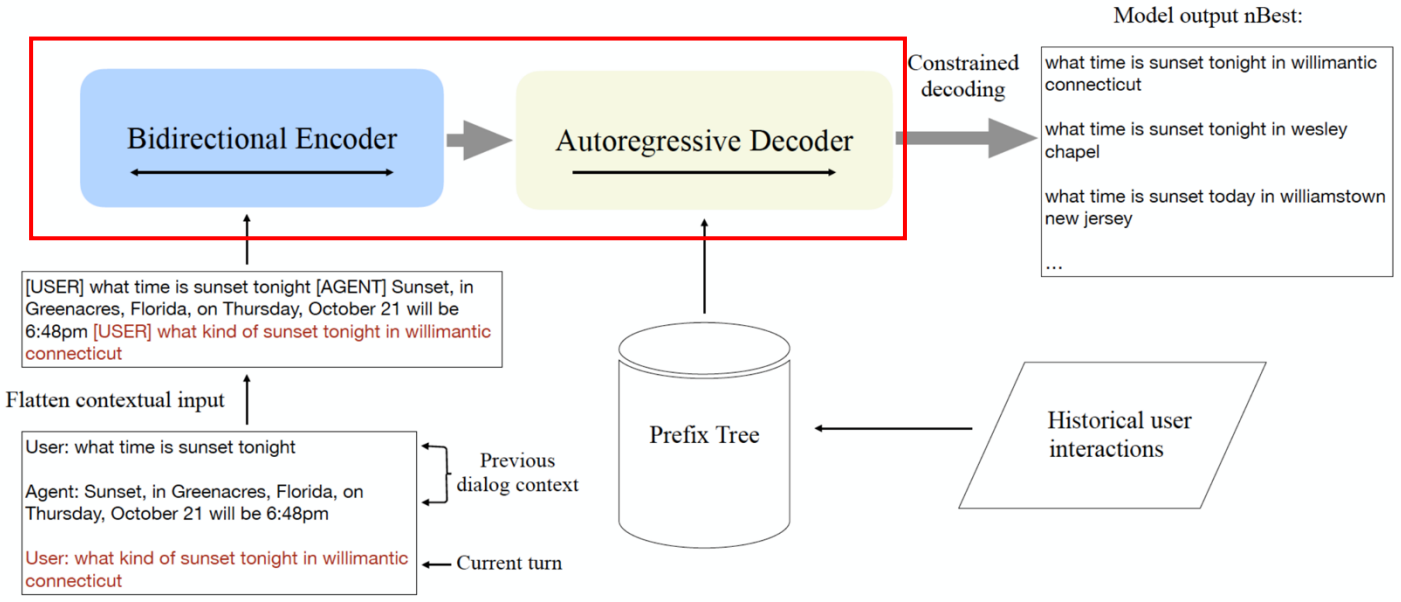


[1] Jie Hao, Yang Liu, Xing Fan, Saurabh Gupta, Saleh Soltan, Rakesh Chada, Pradeep Natarajan, Chenlei Guo, and Gökhan Tür. 2022. CGF: Constrained generation framework for query rewriting in conversational AI. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. 475–483.

Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models
- CGF(EMNLP'22)
  - **1. Encoder–Decoder Architecture**
  - The encoder takes the current query and dialogue history to produce a context-aware representation.
  - The decoder generates the rewritten query.

$$\mathbf{H}_{Enc} = \mathrm{ENC}_{BART}(\mathbf{Q}^0),$$
$$\mathbf{H}_{Dec} = \mathrm{DEC}_{BART}(\mathbf{R}^0, \mathbf{H}_{Enc})$$
$$\mathbf{P}(\mathbf{r_n}|\mathbf{R_{<n}}, \mathbf{Q}; \theta) = Softmax(Proj(\mathbf{h_n}))$$

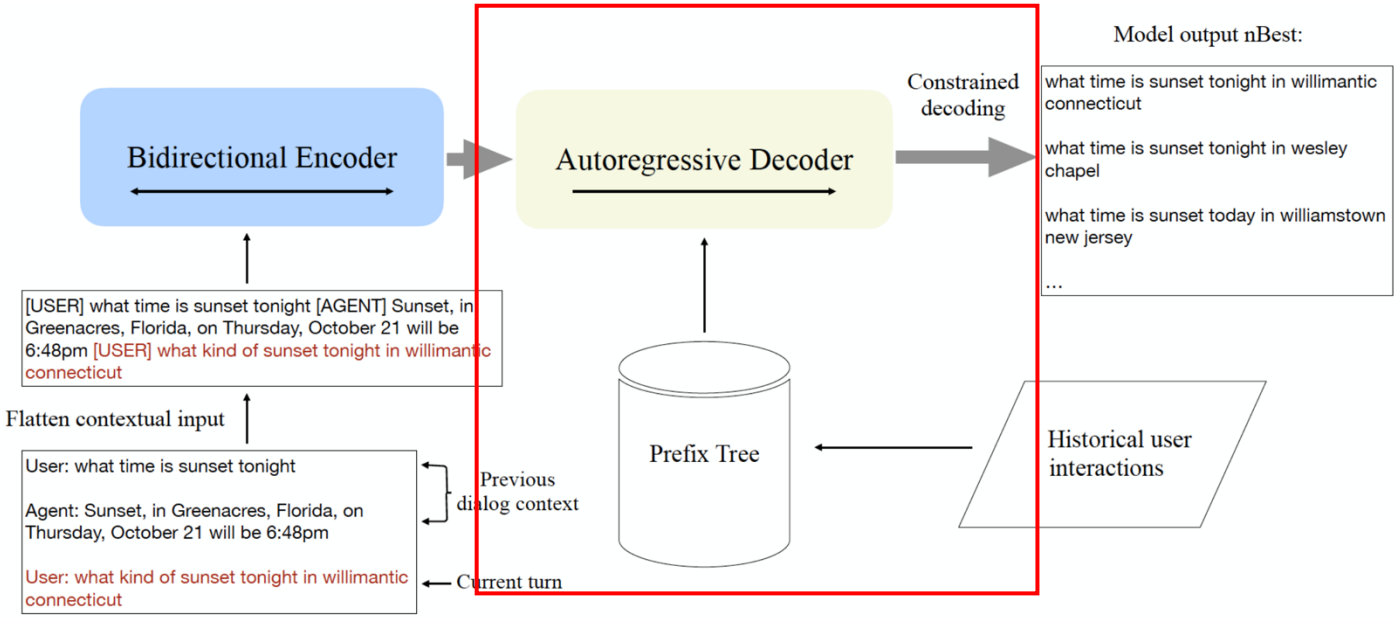


[1] Jie Hao, Yang Liu, Xing Fan, Saurabh Gupta, Saleh Soltan, Rakesh Chada, Pradeep Natarajan, Chenlei Guo, and Gökhan Tür. 2022. CGF: Constrained generation framework for query rewriting in conversational AI. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. 475–483.

Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models
- CGF(EMNLP'22)
  - **2. Constrained Decoding**
  - Instead of freely generating tokens, CGF restricts decoding to a predefined constraint space

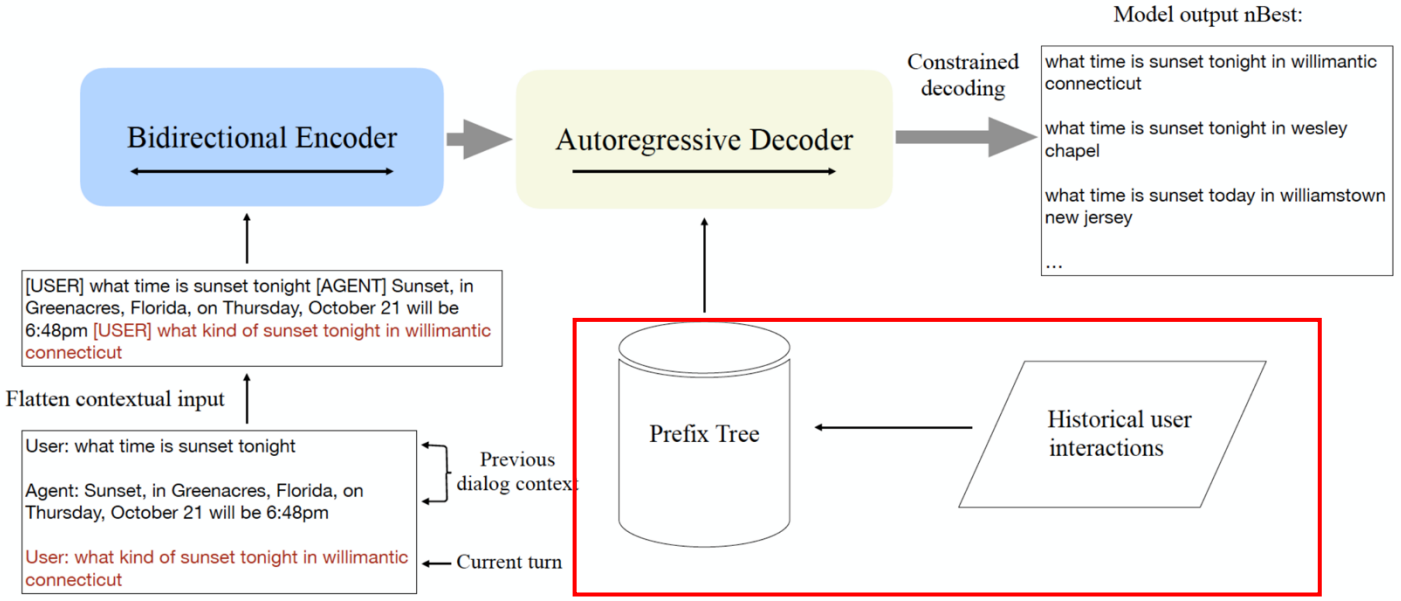
$$\widetilde{P} = \begin{cases} P(\hat{r}_t = r | \hat{r}_{<t}, Q), & \text{if } r \in \text{suffix}_T(\hat{r}_{<t}) \\ 0, & \text{otherwise} \end{cases}$$



[1] Jie Hao, Yang Liu, Xing Fan, Saurabh Gupta, Saleh Soltan, Rakesh Chada, Pradeep Natarajan, Chenlei Guo, and Gökhan Tür. 2022. CGF: Constrained generation framework for query rewriting in conversational AI. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. 475–483.

Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models
- CGF(EMNLP'22)
  - **3. Two Types of Constraints**
  - 1) **Global Constraints**: Derived from common query patterns or corpus-level statistics. Ensure syntactic correctness and general validity
  - 2) **Personalized Constraints**: Derived from user history and preferences. Enable user-specific query rewriting



[1] Jie Hao, Yang Liu, Xing Fan, Saurabh Gupta, Saleh Soltan, Rakesh Chada, Pradeep Natarajan, Chenlei Guo, and Gökhan Tür. 2022. CGF: Constrained generation framework for query rewriting in conversational AI. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. 475–483.

Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models

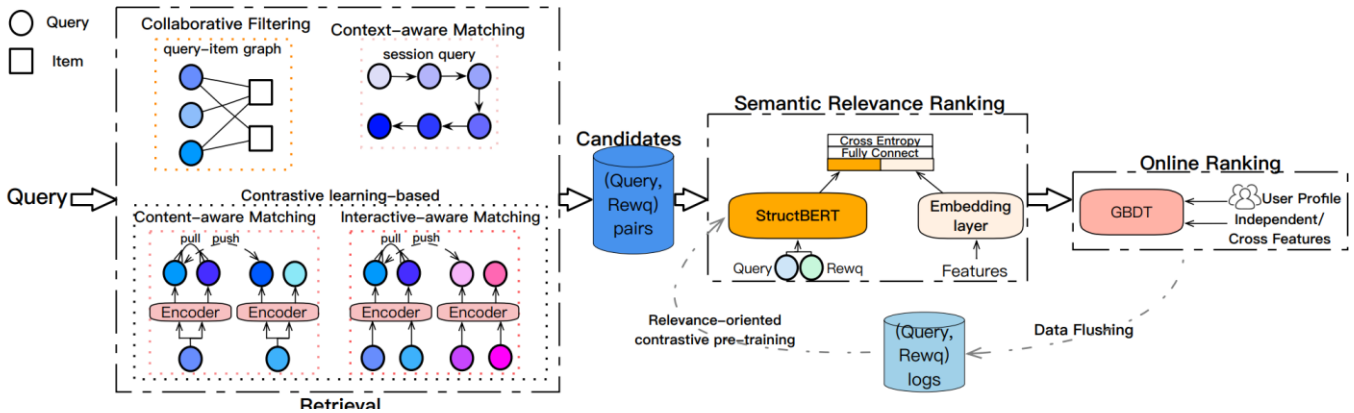
- CGF(EMNLP'22)
  - **CGF rewrites noisy user requests** by correcting spelling errors, speech-recognition mistakes, and ambiguous expressions.
  - **The rewritten queries better capture the user's true intent**, enabling the agent to identify the correct entity, song, or context-dependent request.
  - After rewriting, **agent responses become more accurate**, with defect reductions of 50.20%, 41.52%, and 22.71% in the three examples.

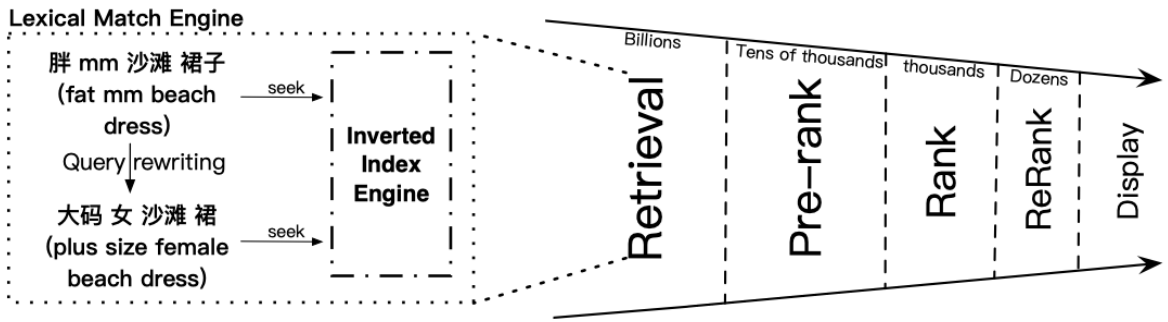
Original request (w/o context): USER: how old is tommy in it **CGF rewrite:** how old is tommyinnit	**Agent response (before rewrite):** Here's what I found from the web: A 24-year-old man whose... **Agent response (after rewrite):** TommyInnit is 17 years old.	**Defect Reduction:** 50.20%
Original request (w/o context): USER: play night talk by drake **CGF rewrite:** play knife talk by drake	**Agent response (before rewrite):** I couldn't find night talk **Agent response (after rewrite):** Here's Knife Talk , by Drake (feat. 21 Savage, Project Pat) , on Amazon Music.	**Defect Reduction:** 41.52%
Original request (w/ context): USER: play little yancy AGENT: Lil' Fancy from Apple Music.y USER: play little yankees praise part **CGF rewrite:** play little yancy praise party	**Agent response (before rewrite):** Sorry, I'm having trouble finding the song. **Agent response (after rewrite):** Here is little yancy praise party.	**Defect Reduction:** 22.71%

[1] Jie Hao, Yang Liu, Xing Fan, Saurabh Gupta, Saleh Soltan, Rakesh Chada, Pradeep Natarajan, Chenlei Guo, and Gökhan Tür. 2022. CGF: Constrained generation framework for query rewriting in conversational AI. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. 475–483.

Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models
- CLE-QR(CIKM'22)
  - **Problem:**
    - Existing QR method ignore personalization leading to degraded search relevance
  - **QR method Require**
    - Reliable generation or large-scale query logs with user feedback

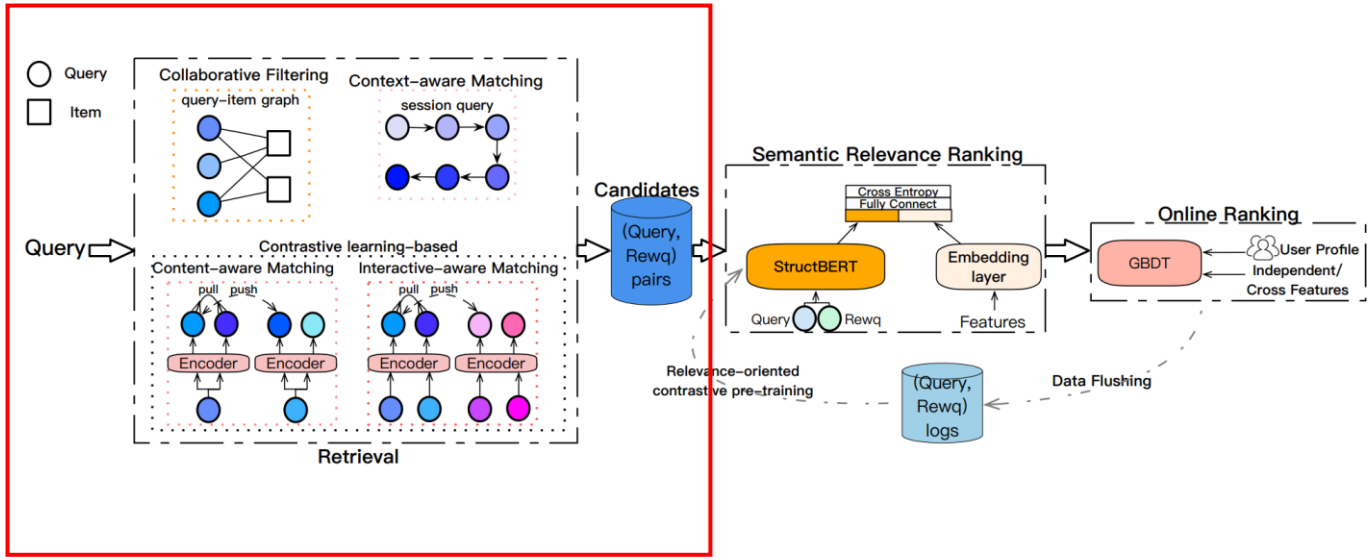




[1] Li, Sen, et al. "Query rewriting in taobao search." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

Direct Personalized Query Rewriting

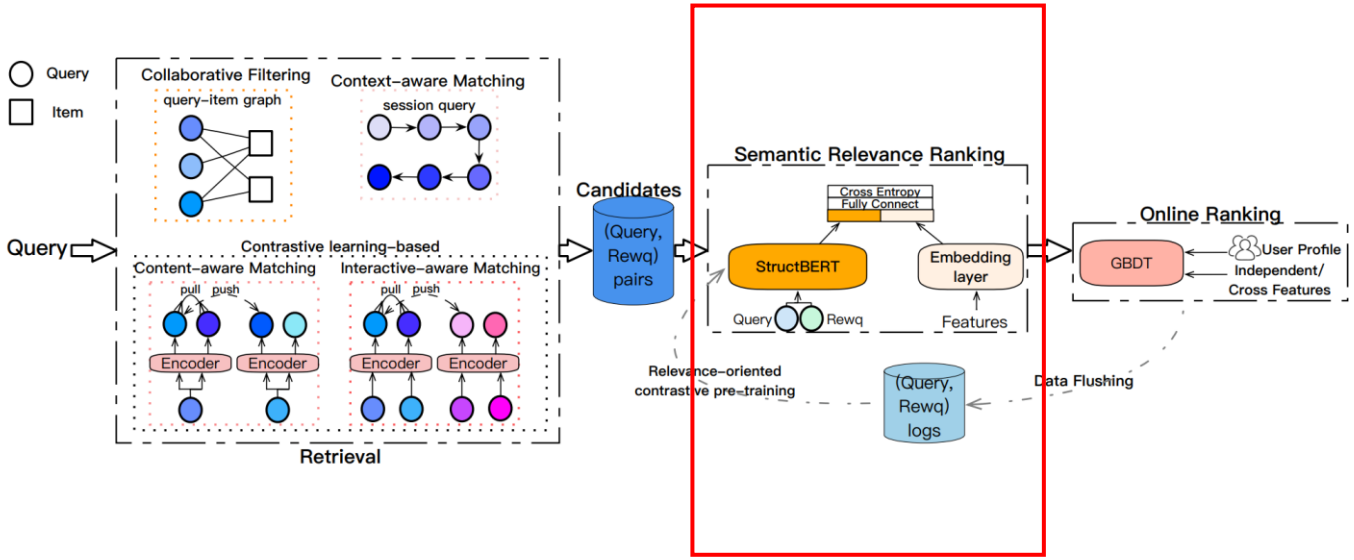
- Focusing on personalized query rewriting by using direct models
- CLE-QR(CIKM'22)
  - **1. Query Retrieval**
    - Retrieve candidate rewrites from large-scale query logs using contrastive representation
      - Collaborative Filtering
      - Embedding-based Retrieval
        - Context-aware Matching
        - Content-aware Matching
        - Interaction-aware Matching



[1] Li, Sen, et al. "Query rewriting in taobao search." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models
- CLE-QR(CIKM'22)
  - **2. Semantic Relevance Ranking**
    - 1)Interaction-aware matching: Select incremental and useful rewrites
    - 2)Contrastive learning with user feedback. Improves discrimination between good/bad rewrites



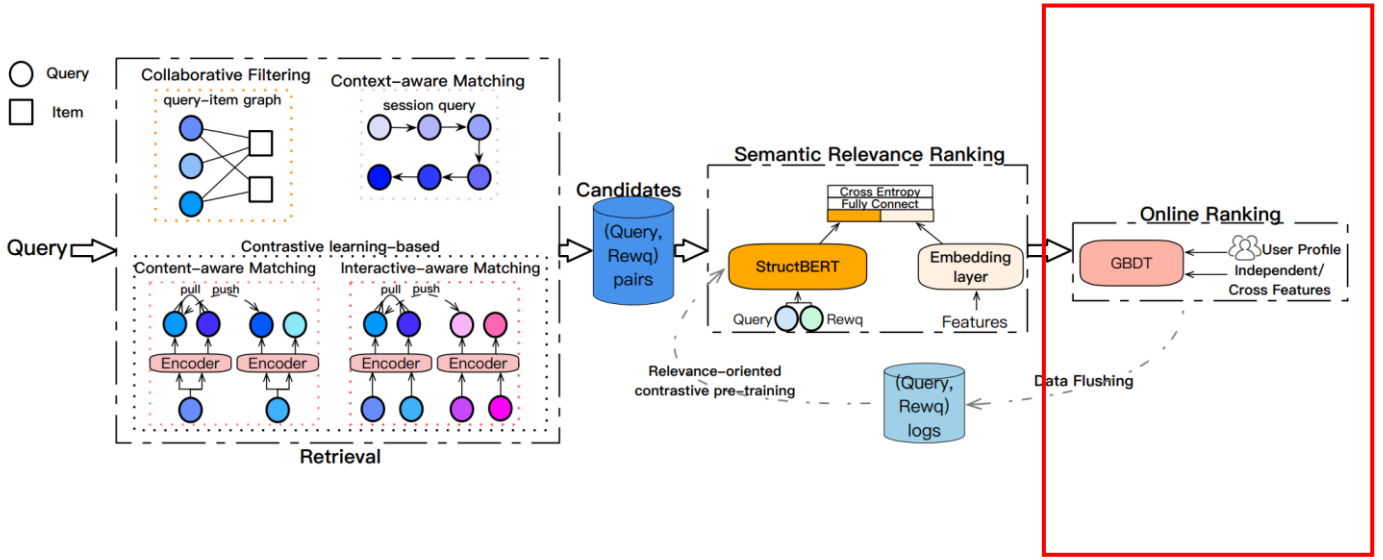
$$e_{f_i} = W_{f_i} \cdot f_i,$$
$$h_f = concat(\{e_{f_i} | f_i \in \mathcal{F}\}),$$
$$h_s = Encoder^{first}([CLS]Query[SEP]Rewq[SEP]),$$
$$h = W_h[h_f, h_s],$$

$$\ell = -\frac{1}{N}\sum_i [y_i \cdot \log(\sigma(h_i)) + (1 - y_i) \cdot \log(1 - \sigma(h_i))],$$

[1] Li, Sen, et al. "Query rewriting in taobao search." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models
- CLE-QR(CIKM'22)
  - **3. Online Ranking**
    - Final ranking incorporates user profile and behavior signals



[1] Li, Sen, et al. "Query rewriting in taobao search." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

Direct Personalized Query Rewriting

- Focusing on personalized query rewriting by using direct models

- CLE-QR(CIKM'22)
  - **Collaborative filtering tends to retireval behavior-related** but sometimes noisy rewrites, such as brand names or loosely related products that may not fully preserve the original query intent.
  - **Context2vec can effectively reduce the semantic gap** and focus more on semantic matching than Collaborative Filtering.
  - **Interaction-aware rewriting further incorporates user behavior signals,** producing rewrites that are closer to real search and click patterns.

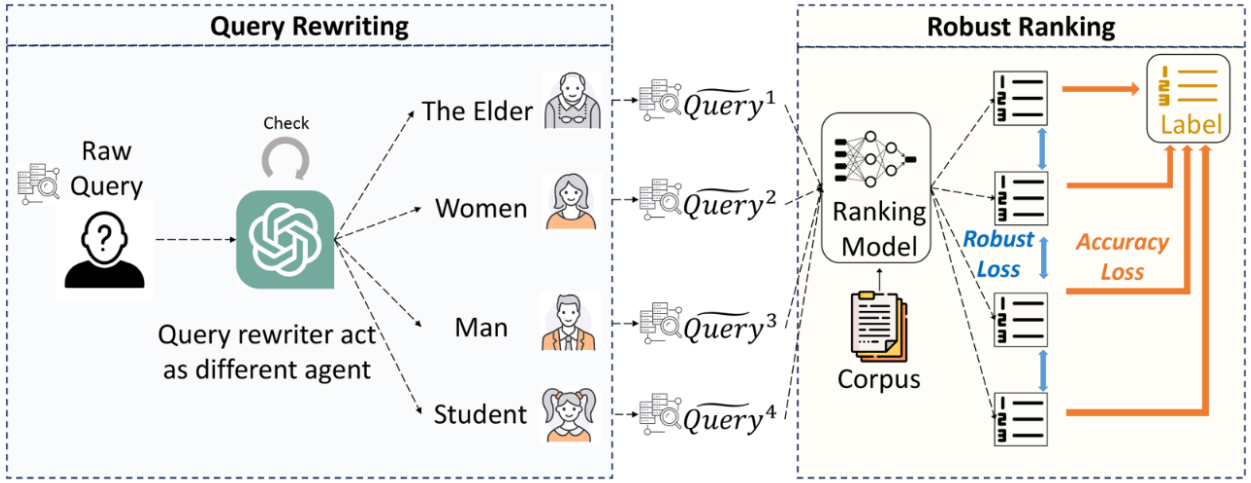
Original Query	*Collaborative Filtering*	*Context2vec*	*Content2vec*	*Content-aware*	*Interaction-aware*
水果味薄荷糖 (fruit flavored mints)	亲劲 (a mint brand)	有个圈的薄荷糖 (mints with a circle)	口香糖 (chewing gum)	水果薄荷糖 (fruit mints)	果味糖 (-)
	果味软糖 (fruity fudge)	mints (mints)	mints海盐薄荷糖 (sea salt mints)	无糖水果薄荷糖 (sugar-free fruit mints)	水果糖 (fruit candy)
	益美滋 (a mint brand)	水蜜桃薄荷糖 (peach mints)	提神薄荷糖 (refreshing mints)	百香果薄荷糖 (passion Fruit Mints)	水果薄荷糖 (fruit mints)
jordan1 童鞋 (jordan1 kids shoes)	6351 ar (nike shoes model)	6066 bq (nike shoes model)	耐克空军一号 (nike air force 1)	jordan 童鞋 (jordan kids shoes)	aj1 童鞋 (aj1 kids shoes)
	nike aj 童鞋 (nike aj kids shoes)	耐克aj儿童 (nike aj kids)	aj14 (aj14)	jordan11 童鞋 (jordan11 kids shoes)	aj儿童鞋 (aj kids shoes)
	6066 bq (nike shoes model)	aj13童鞋 (aj13 kids shoes)	耐克儿童鞋子 (nike kids shoes)	airforce 1 童鞋 (airforce1 kids shoes)	乔丹童鞋 (jordan kids shoes)

[1] Li, Sen, et al. "Query rewriting in taobao search." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.

Auxiliary Personalized Query Rewrite

- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- Agent4Ranking (Our method)

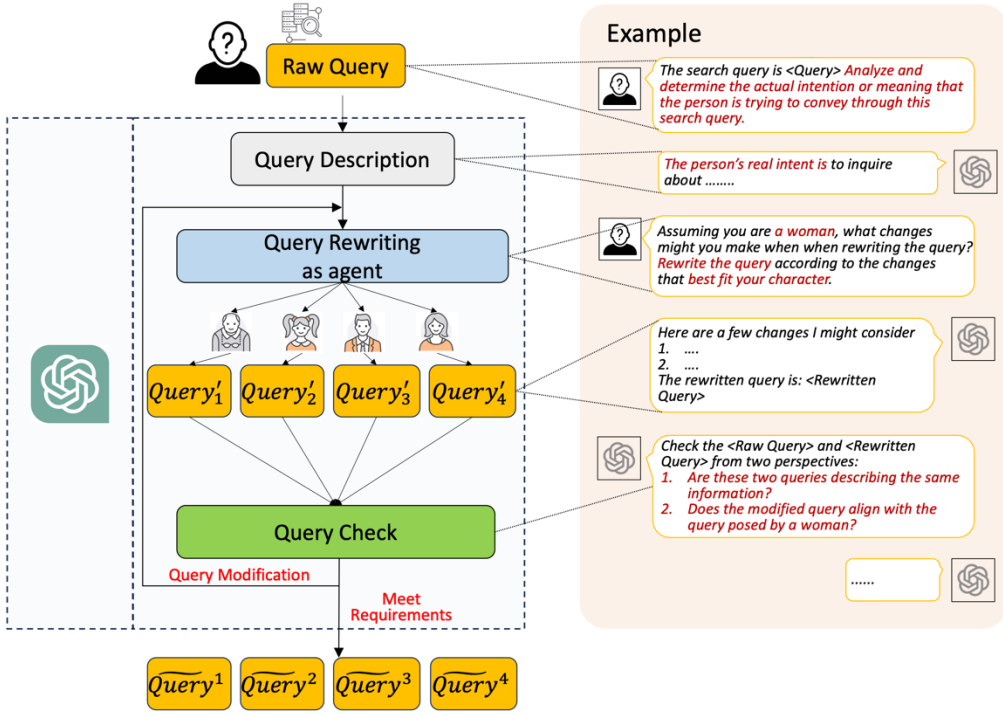
  - **Problem:**
    - Traditional ranking systems assume that: semantically equivalent queries should yield similar ranking results.
    - However, Users from different demographics (e.g., age, background, habits) express the same intent **differently**
  - **QR method Require**
    - Incorporate demographic diversity in query formulation



[1] Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. 2023. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm. arXiv preprint arXiv:2312.15450 (2023).

Auxiliary Personalized Query Rewrite

- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- Agent4Ranking (Our method)
  - **1. Multi-Agent Query Rewriting**
  - Use multiple LLM agents, each representing a different user persona.
  - Each agent is prompted with a demographic profile and rewrites the same query accordingly

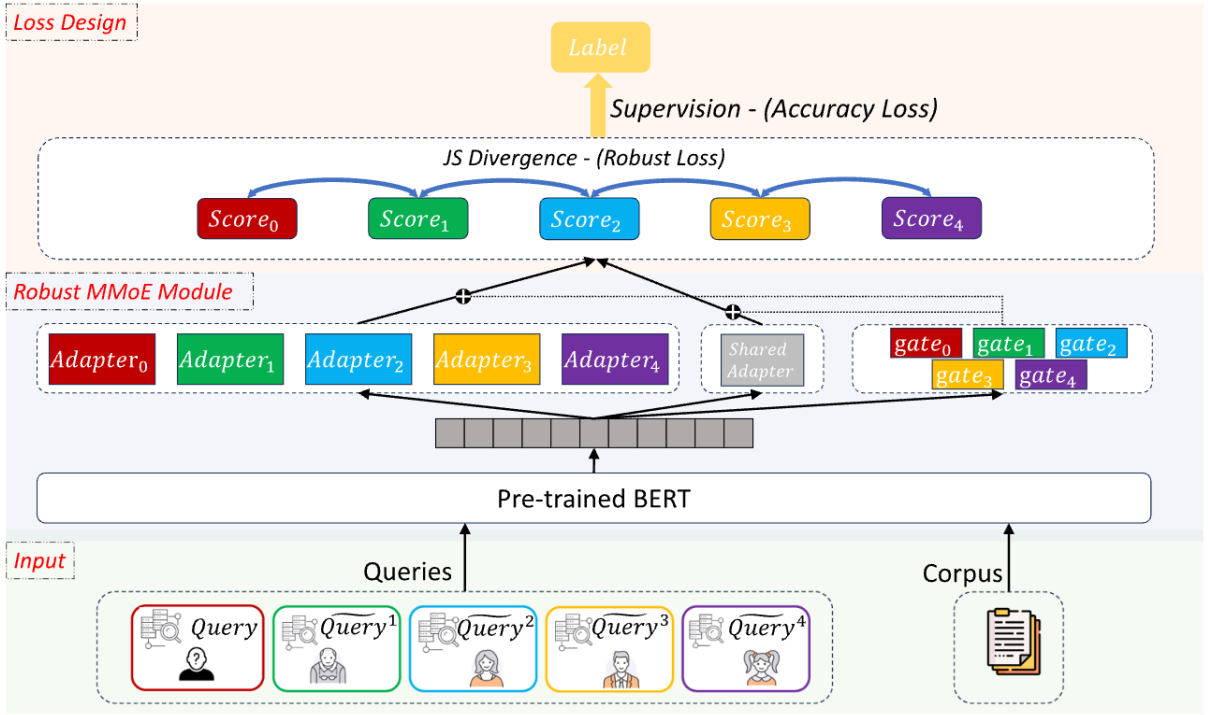


[1] Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. 2023. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm. arXiv preprint arXiv:2312.15450 (2023).

Auxiliary Personalized Query Rewrite

- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- Agent4Ranking (Our method)

  - **2. Robust Ranking Model**
  - 1)**Query-Specific Experts**: Each rewritten query (from different agents/personas) has its own expert. Captures stylistic variation and demographic-specific expression
  - 2)**Query-Shared Expert:** A shared expert across all rewrites. Captures invariant semantic intent and core query meaning



$$g_i^k = \text{softmax}(W^k e_i^k)$$

$$h_i^k = g_i^k [v_i^k, w_i^k]$$

$$\hat{y}_i^k = \mathcal{C}(h_i^k)$$

[1] Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. 2023. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm. arXiv preprint arXiv:2312.15450 (2023).

Auxiliary Personalized Query Rewrite

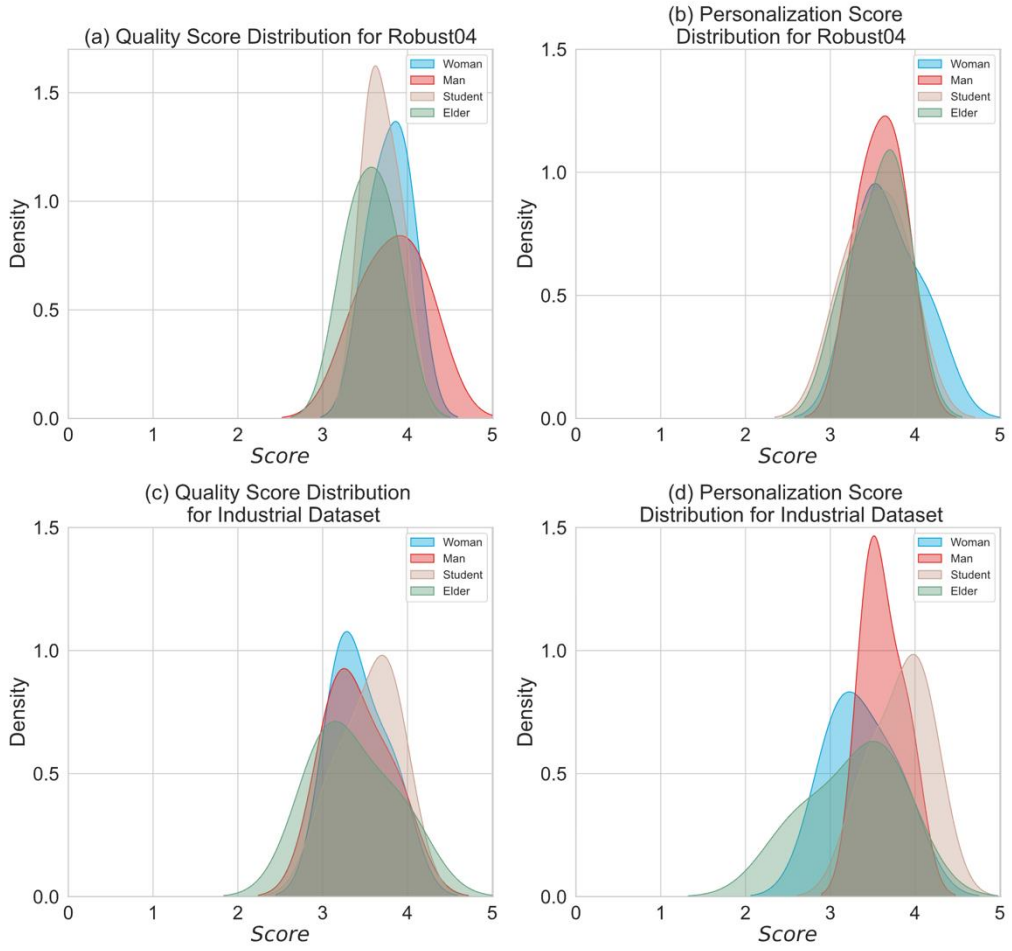
- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- Agent4Ranking (Our method)
  - **For example, the rewriting of different roles usingLLMs exhibits distinct characteristics.**

Public Dataset:		Industrial Dataset:	
Case 1:		***Case 1:***	
Original:	Has prayer in U.S. schools been banned completely?	*Original:*	眼角膜烧伤恢复得多长时间
Woman:	Is it true that prayer has been completely banned in schools across the United States?	*Woman:*	眼角膜烧伤需要多久才能康复
Man:	Has prayer been completely banned in schools in the U.S.?	*Man:*	眼角膜烧伤得多长时间恢复
Old:	Are prayers still permitted in schools in the United States?	*Old:*	眼角膜烧伤需要多久才能完全恢复?
Student:	Is prayer in U.S. schools completely banned?	*Student:*	眼角膜烧伤恢复时间
Case 2:		***Case 2:***	
Original:	How are Oscar winners selected?	*Original:*	如何制作一道蓝莓山药泥
Woman:	How does the selection process for Oscar winners work?	*Woman:*	如何做蓝莓山药泥
Man:	How are the winners of the Oscars determined?	*Man:*	蓝莓山药泥如何制作
Old:	How do they decide who wins an Oscar?	*Old:*	怎样才能做出蓝莓山药
Student:	What's the process for selecting Oscar winners?	*Student:*	蓝莓山药泥的做法

[1] Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. 2023. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm. arXiv preprint arXiv:2312.15450 (2023).

Auxiliary Personalized Query Rewrite

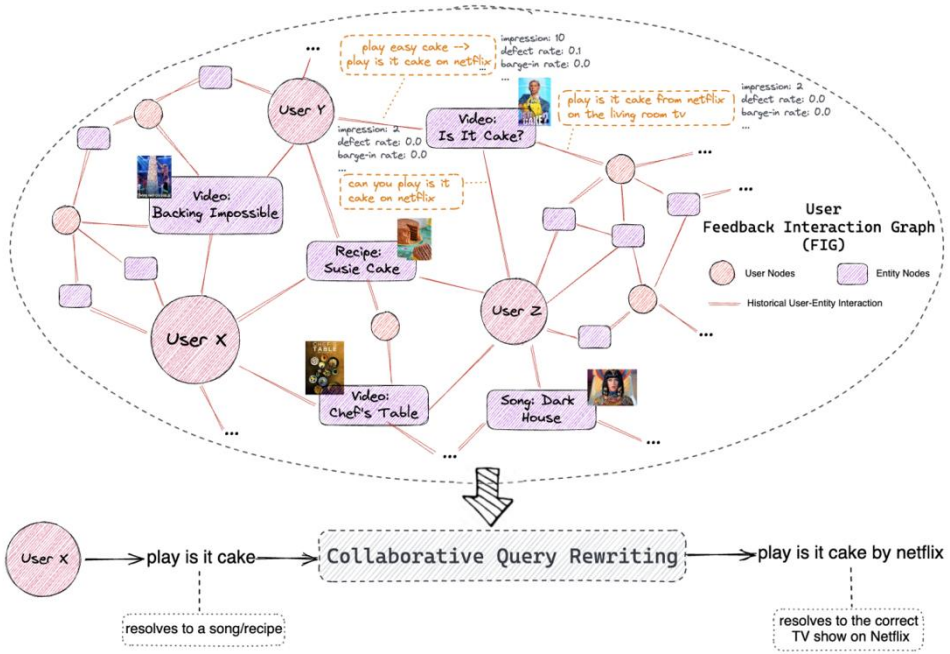
- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- Agent4Ranking (Our method)
  - They evaluates the quality of rewritten queries across different agent roles, including Woman, Man, Student, and Elder.
  - Both query quality and personalization scores are concentrated around high values, mostly between 3 and 4.5 on a 5-point scale.
  - The score distributions are relatively stable across roles, suggesting that the method does not strongly favor a specific role type.



[1] Xiaopeng Li, Lixin Su, Pengyue Jia, Xiangyu Zhao, Suqi Cheng, Junfeng Wang, and Dawei Yin. 2023. Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm. arXiv preprint arXiv:2312.15450 (2023).

Auxiliary Personalized Query Rewrite

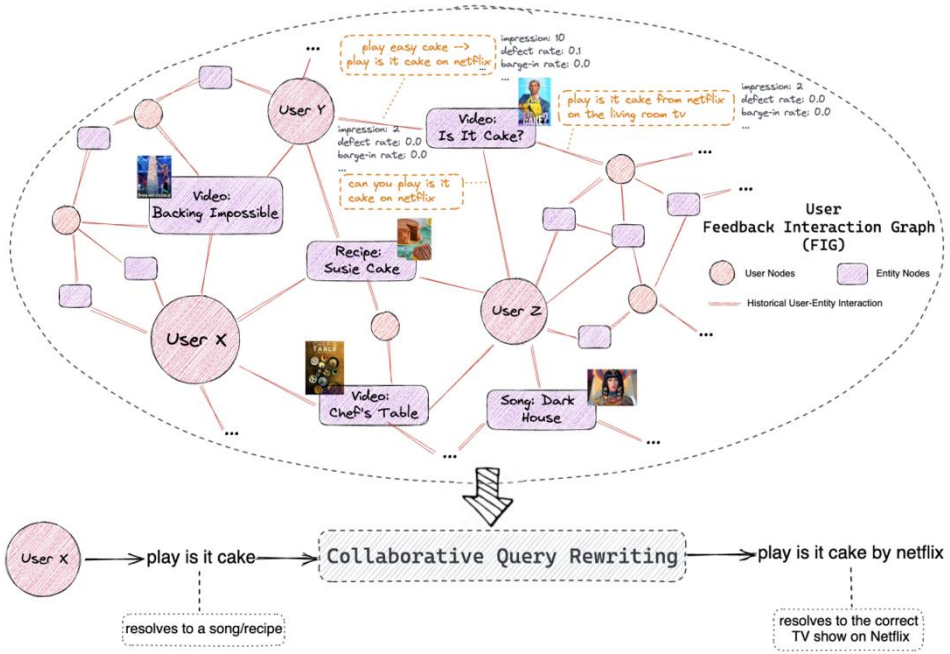
- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- FIG (EMNLP 2023)
  - **Problem:**
    - Existing methods fails to handle unseen user queries beyond historical interactions
  - **QR method Require**
    - Expand from individual history to collaborative knowledge



[1]] Chen, Zheng, et al. "Graph Meets LLM: A Novel Approach to Collaborative Filtering for Robust Conversational Understanding." EMNLP: Industry Track. 2023.

Auxiliary Personalized Query Rewrite

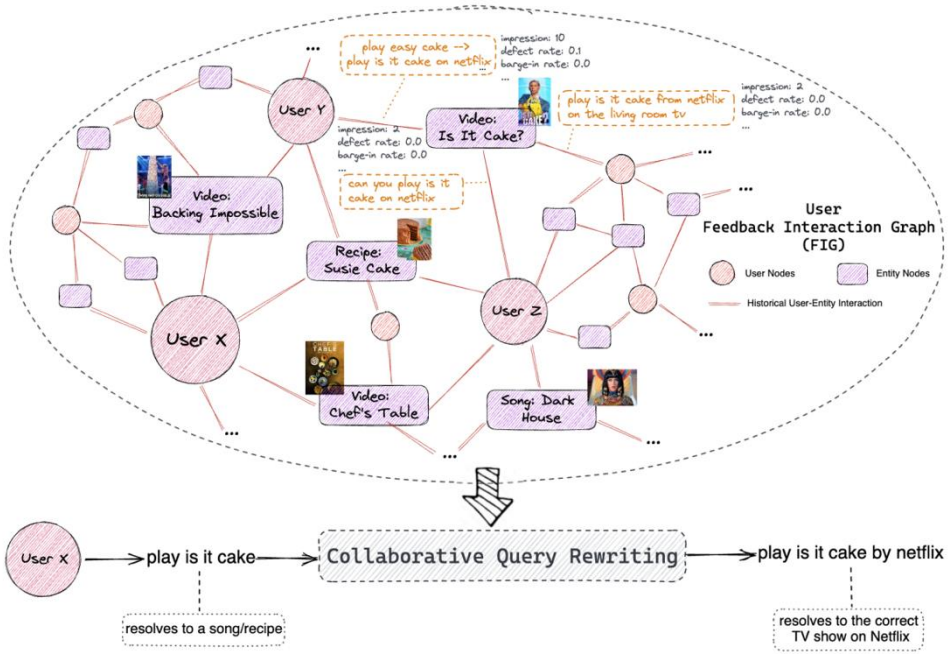
- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- FIG (EMNLP 2023)
  - **1. User Feedback Interaction Graph (FIG)**
    - Build graph
      - nodes: users / entities
      - edges: interaction history



[1]] Chen, Zheng, et al. "Graph Meets LLM: A Novel Approach to Collaborative Filtering for Robust Conversational Understanding." EMNLP: Industry Track. 2023.

Auxiliary Personalized Query Rewrite

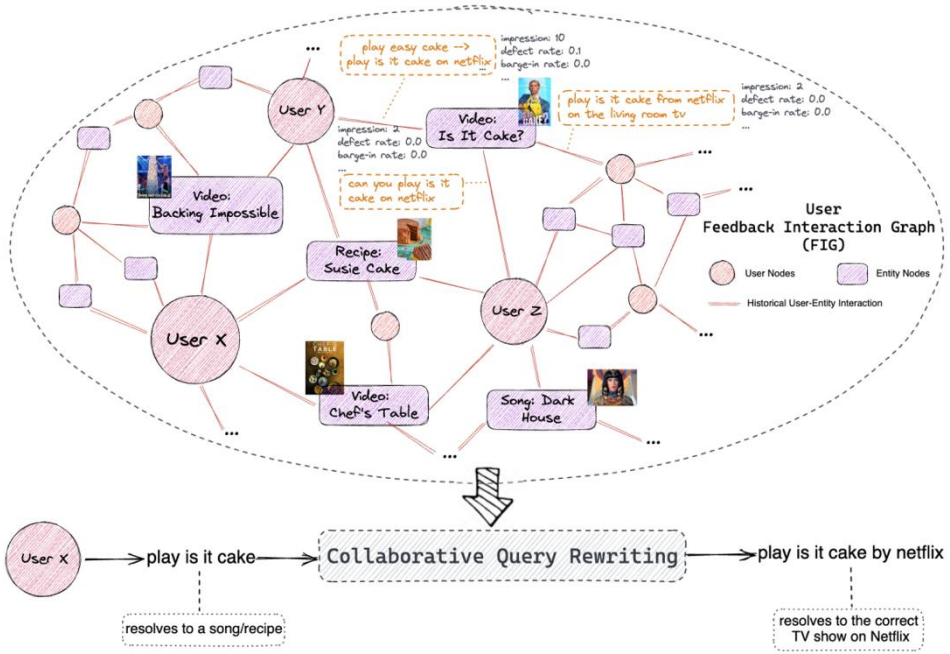
- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- FIG (EMNLP 2023)
  - **2. Multi-hop Graph Traversal**
    - Traverse graph to:
    - discover related entities beyond user history
    - Generate new candidate interactions
    - Solves unseen query problem



[1]] Chen, Zheng, et al. "Graph Meets LLM: A Novel Approach to Collaborative Filtering for Robust Conversational Understanding." EMNLP: Industry Track. 2023.

Auxiliary Personalized Query Rewrite

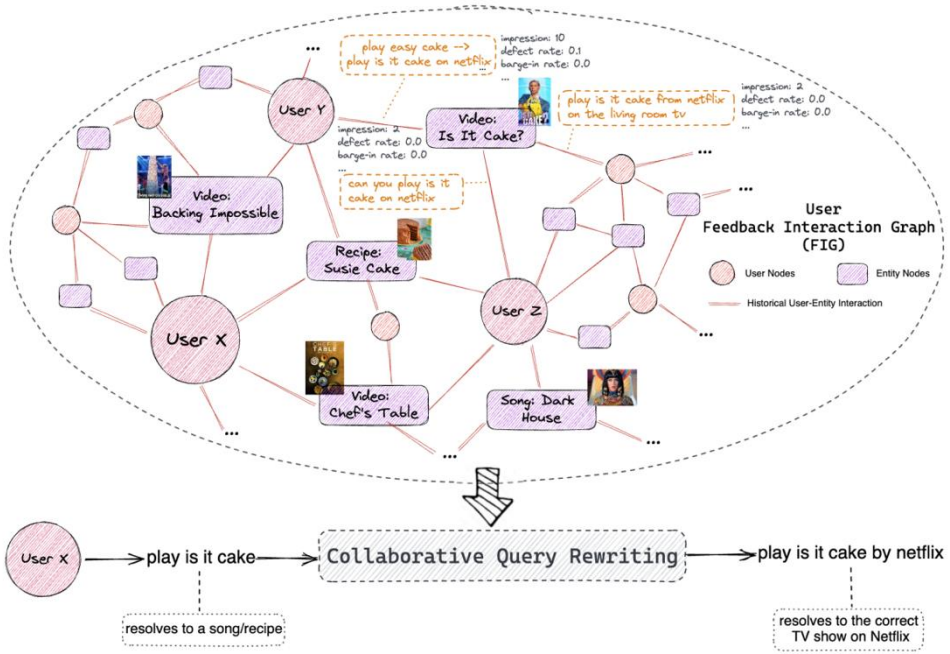
- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- FIG (EMNLP 2023)
  - **3. Collaborative User Index**
    - Expand original user index by adding graph-inferred entries
    - Result in broader coverage for future queries



[1]] Chen, Zheng, et al. "Graph Meets LLM: A Novel Approach to Collaborative Filtering for Robust Conversational Understanding." EMNLP: Industry Track. 2023.

Auxiliary Personalized Query Rewrite

- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- FIG (EMNLP 2023)
  - **4. LLM Enhancement**
    - Use LLM (e.g., Dolly-V2) to predict new user–entity links for augment graph further improves coverage



[1]] Chen, Zheng, et al. "Graph Meets LLM: A Novel Approach to Collaborative Filtering for Robust Conversational Understanding." EMNLP: Industry Track. 2023.

Auxiliary Personalized Query Rewrite

- Emphasizes personalized query rewriting by using auxiliary mechanisms, such as retrieval, reasoning strategies, and external memory.
- FIG (EMNLP 2023)
  - Graph traversal alone provides limited unseen user interaction coverage, with relatively low coverage on both video and music domains.
  - Adding Dolly-V2 link prediction improves coverage, especially after fine-tuning.
  - Fine-tuned Dolly-V2 achieves the best performance, increasing coverage to 24.5% for video and 18.4% for music when the index size cap is 200.
  - The results show that LLM-based link prediction can better discover potential user-item connections, improving the effectiveness of collaborative user indexes.

	Video			Music		
Index Construction Method	**100**	**200**	**500**	**100**	**200**	**500**
Graph Traversal Only	1.8%	3.8%	6.3%	1.1%	2.7%	5.4%
+Dolly-V2 Link Prediction (not fine-tuned)	2.5%	5.3%	N/A	1.4%	3.6%	N/A
+Dolly-V2 Link Prediction (fine-tuned)	**10.8%**	**24.5%**	N/A	**8.5%**	**18.4%**	N/A

Table 2: Comparison of unseen user interaction coverage by collaborative user indexes constructed by different methods, with index size cap 100, 200 and 500.

[1]] Chen, Zheng, et al. "Graph Meets LLM: A Novel Approach to Collaborative Filtering for Robust Conversational Understanding." EMNLP: Industry Track. 2023.

Pre-retrieval: Query Rewrite Overview

Method	Category	Paper
Cho, Eunah, et al	Direct Model-based	Cho, Eunah, et al. "Personalized search-based query rewrite system for conversational ai." Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI. 2021.
Nguyen, Duy A., et al	Direct Model-based	Nguyen, Duy A., et al. "RL-based Query Rewriting with Distilled LLM for online E-Commerce Systems." arXiv preprint arXiv:2501.18056 (2025).
Li, Sen, et al	Direct Model-based	Li, Sen, et al. "Query rewriting in taobao search." Proceedings of the 31st ACM International Conference on Information & Knowledge Management. 2022.
CGF	Direct Model-based	Hao, Jie, et al. "CGF: Constrained generation framework for query rewriting in conversational AI." Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. 2022.
Li, Cheng, et al	Direct Model-based	Li, Cheng, et al. "Learning to rewrite prompts for personalized text generation." Proceedings of the ACM Web Conference 2024. 2024.
PEARL	Direct Model-based	Mysore, Sheshera, et al. "Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers." arXiv preprint arXiv:2311.09180 (2023).
Agent4Ranking	Auxiliary-enhanced	Li, Xiaopeng, et al. "Agent4ranking: Semantic robust ranking via personalized query rewriting using multi-agent llm." arXiv preprint arXiv:2312.15450 (2023).
Zhou, Denny, et al	Auxiliary-enhanced	Zhou, Denny, et al. "Least-to-most prompting enables complex reasoning in large language models." arXiv preprint arXiv:2205.10625 (2022).
ERAgent	Auxiliary-enhanced	Shi, Yunxiao, et al. "Eragent: Enhancing retrieval-augmented language models with improved accuracy, efficiency, and personalization." arXiv preprint arXiv:2405.06683 (2024).
Zhou, Yujia, et al	Auxiliary-enhanced	Zhou, Yujia, et al. "Cognitive personalized search integrating large language models with an efficient memory mechanism." Proceedings of the ACM Web Conference 2024. 2024.
Chen, Zheng, et al	Auxiliary-enhanced	Chen, Zheng, et al. "Graph meets LLM: A novel approach to collaborative filtering for robust conversational understanding." arXiv preprint arXiv:2305.14449 (2023).
BASES	Auxiliary-enhanced	Ren, Ruiyang, et al. "Bases: Large-scale web search user simulation with large language model based agents." arXiv preprint arXiv:2402.17505 (2024).

Pre-retrieval: Query Expansion

- Motivation
  - ***Original query fails to capture user-specific intent.***

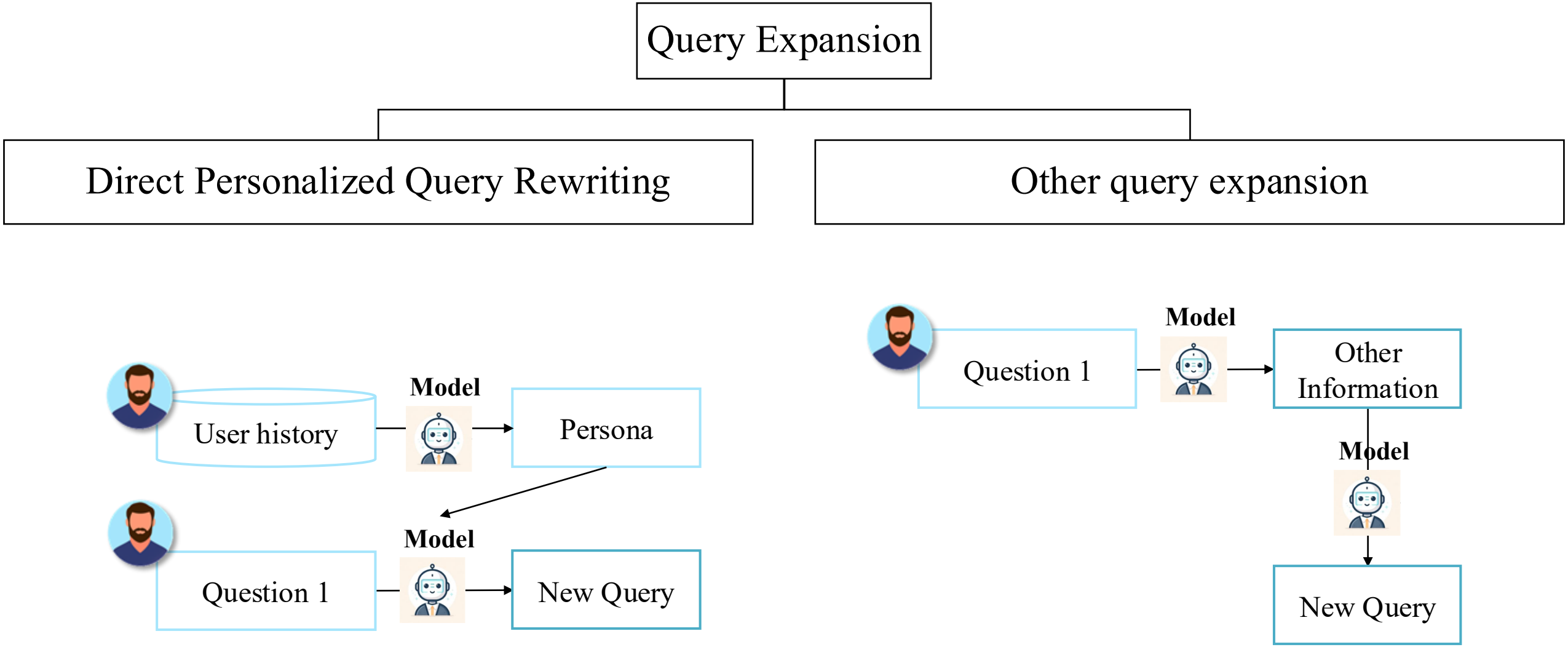
Pre-retrieval: Query Expansion

- Motivation
  - ***Original query fails to capture user-specific intent.***

- Personalized Query Expansion Definition
  - Query expansion enhances retrieval systems by ***expanding a user's original query with additional terms, synonyms, or refined structure***. This improves the relevance and scope of retrieved documents

- Methods can be divided to:
  - Direct Personalized Query Rewriting
  - Others

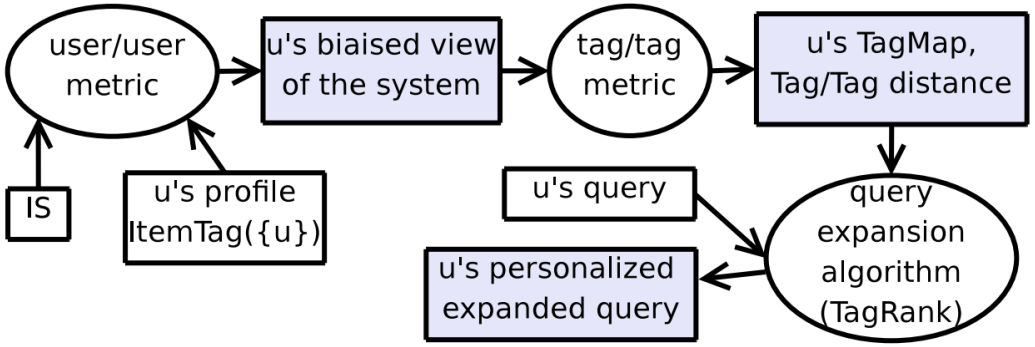
Pre-retrieval: Query Expansion



Direct Personalized Query Expansion

- Incorporating tagging information to enhance personalized query expansion

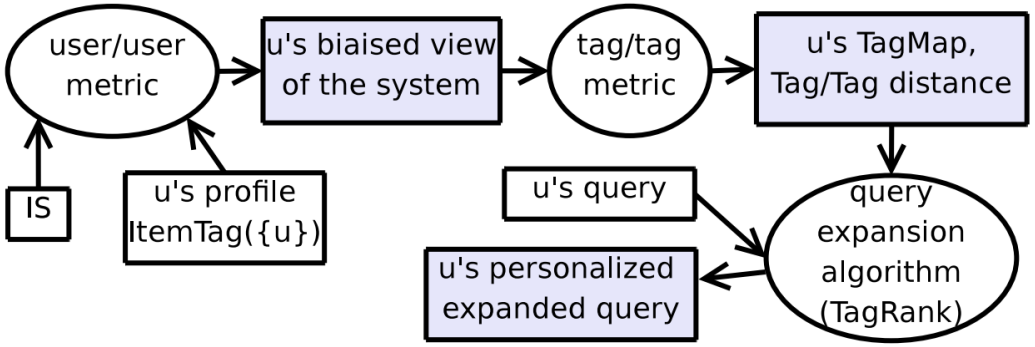
- Gossple
  - **Problem:**
    - Traditional query expansion methods ignore user-specific preferences and rely on global statistics.
    - This leads to the vocabulary mismatch problem
  - **QE method Require**
    - Incorporate users generate personal vocabularies (tags)



[1]] Bertier, Marin, et al. "Toward personalized query expansion." Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. 2009.

Direct Personalized Query Expansion

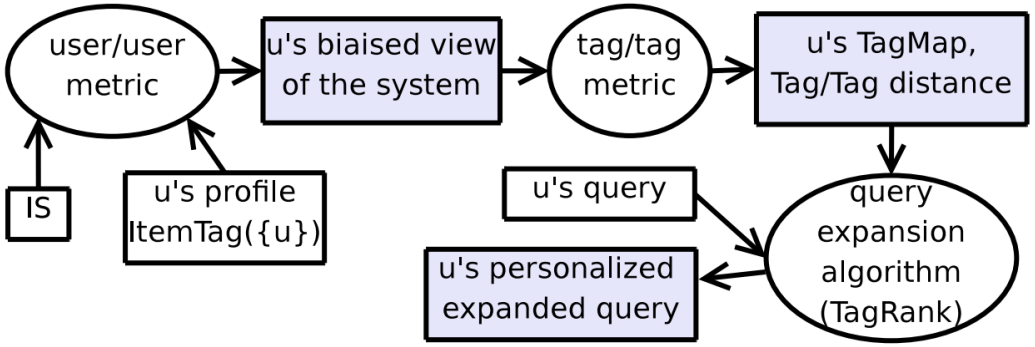
- Incorporating tagging information to enhance personalized query expansion
- Gossple
  - **1. User-Centric Expansion**
    - Instead of selecting expansion terms globally, the model constructs a user profile and identifies similar users.
    - For extracts expansion terms from user's own tags and sneighbors' tags



[1]] Bertier, Marin, et al. "Toward personalized query expansion." Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. 2009.

Direct Personalized Query Expansion

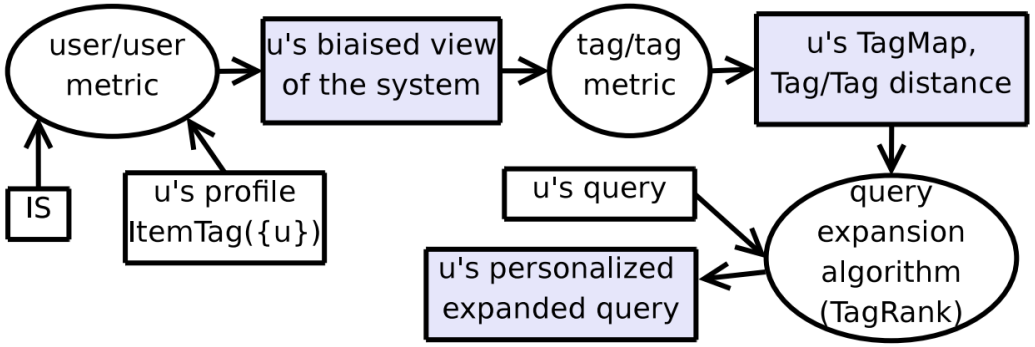
- Incorporating tagging information to enhance personalized query expansion
- Gossple
  - **2. Social Network Modeling**
    - The paper introduces a personalized network:
    - Nodes: users and tags
    - Edges: shared interests and tagging similarity
    - closer users → more influence



[1]] Bertier, Marin, et al. "Toward personalized query expansion." Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. 2009.

Direct Personalized Query Expansion

- Incorporating tagging information to enhance personalized query expansion
- Gossple
  - **3. Expansion Strategy**
    - Expansion terms are selected based on user's own tagging behavior, Tags from similar users, and Semantic relatedness in the tagging space



[1]] Bertier, Marin, et al. "Toward personalized query expansion." Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. 2009.

Direct Personalized Query Expansion

- Personalizes the query before retrieval by incorporating user history into both query expansion and semantic alignment.

- PBR (Ours AAAI-26 Oral)
  - **Problem:**
    - User expression styles are inherently diverse:
    - User corpora induce heterogeneous semantic structures
  - **QE method Require**
    - Align with individual user intent and corpus characteristics both in style and structure

User 1 Memory



It's always fun to prepare, and using **organic ingredients** really makes the difference.

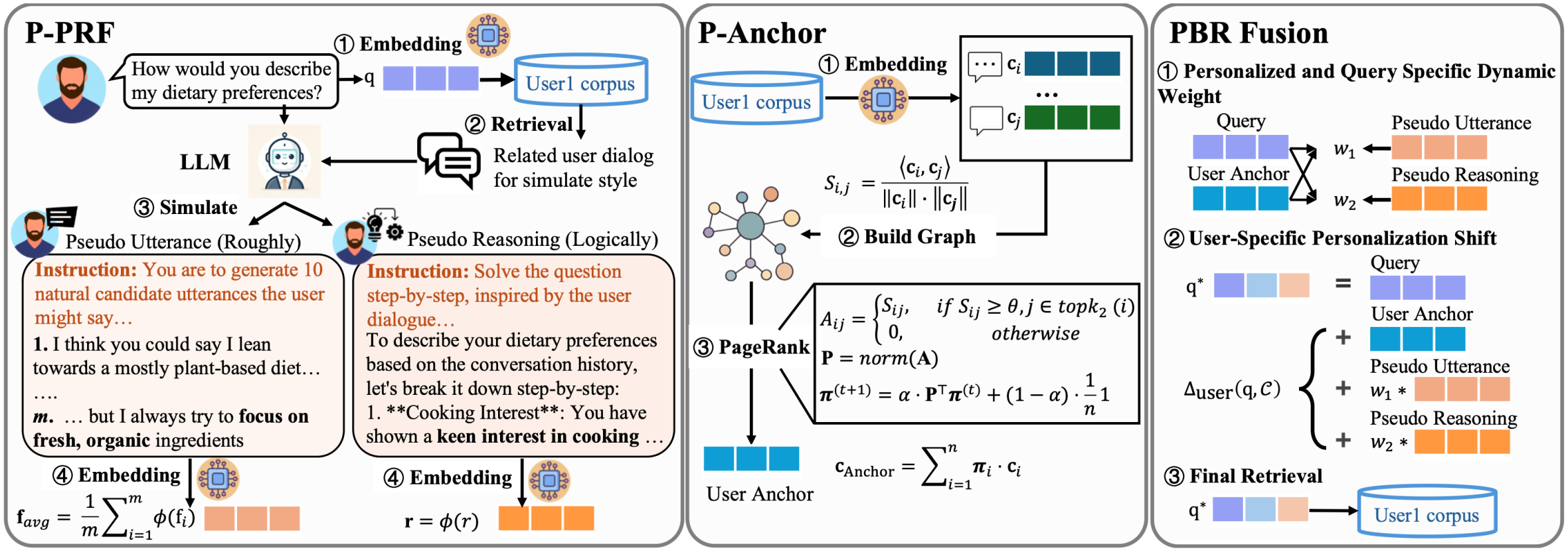
User 2 Memory



We have all the ingredients, and I'm **trying a new side dish recipe** I found too ...

Direct Personalized Query Expansion

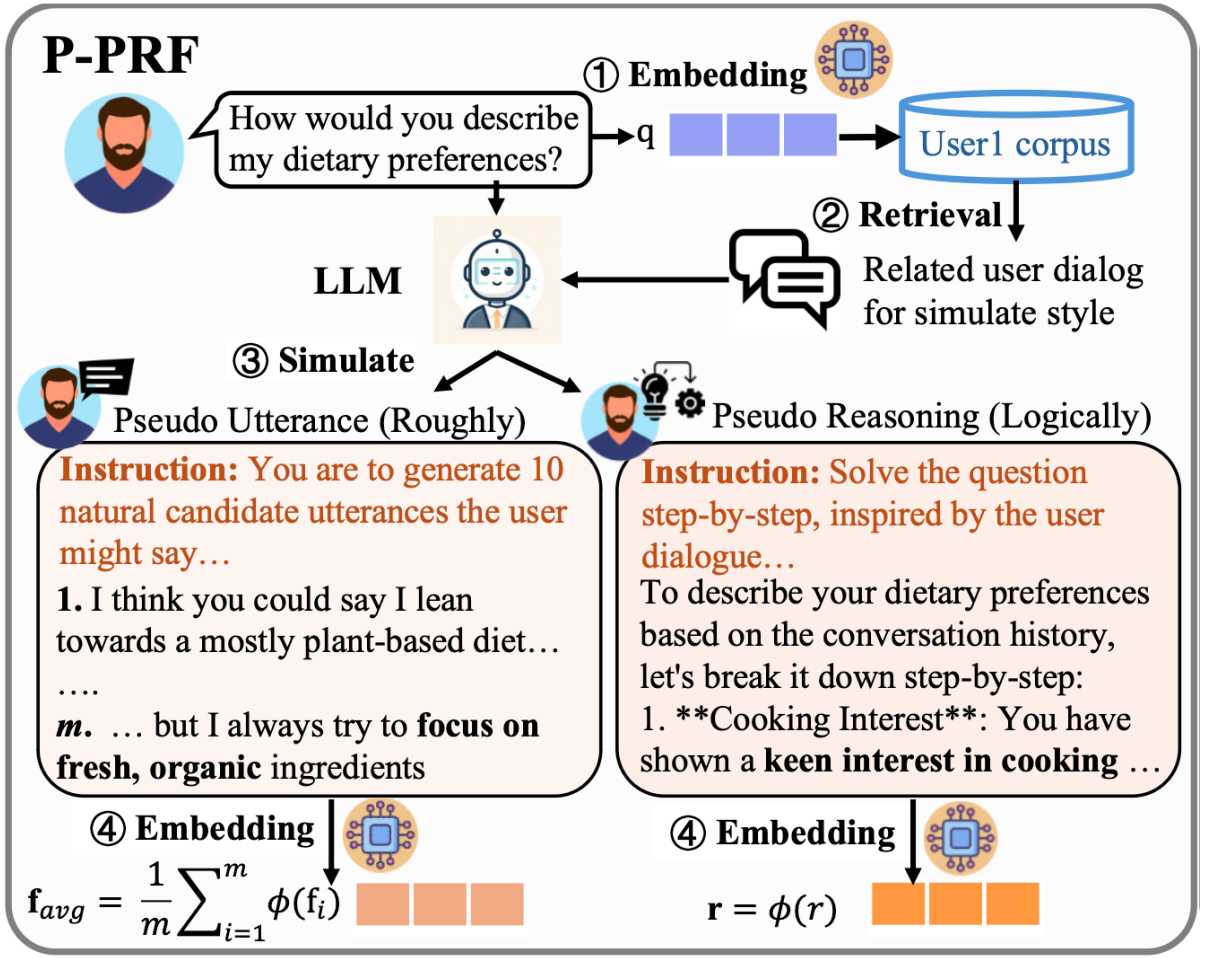
- Personalizes the query before retrieval by incorporating user history into both query expansion and semantic alignment.
- PBR (Ours AAAI-26 Oral)



Direct Personalized Query Expansion

P-PRF: Style-aware question expansion module

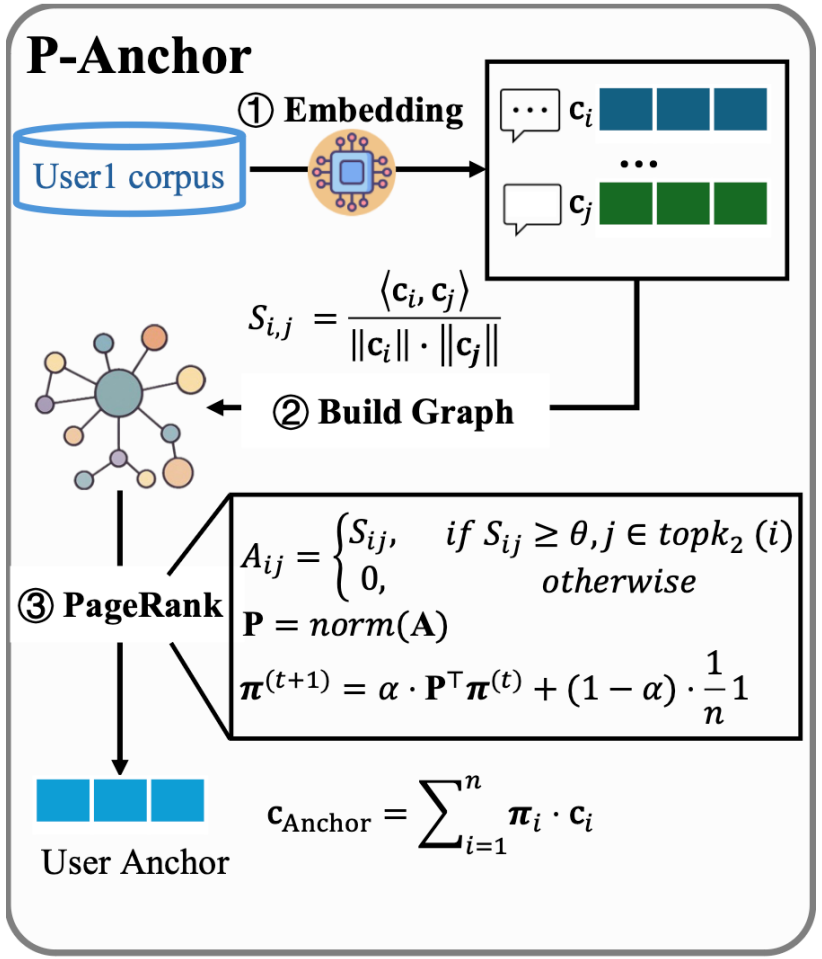
- **Motivation**
  - User have specific expression style, making it difficult to construct expansions that faithfully preserve user-specific semantics
- **Our Idea**
  - We utilizing LLM to simulates *two complementary forms* of feedback to generate pseudo utterances and pseudo reasoning that reflect the user's stylistic tendencies and underlying reasoning logic



Direct Personalized Query Expansion

P-Anchor： Graph-based alignment module

- **Motivation**
  - Existing methods are often designed to retrieve from a unified external corpus, lacking explicit grounding in the structural organization of individual user corpora
- **Our Idea**
  - We captures corpus-level preferences by identifying semantically central regions within a graph-structured user corpus embedding space



Direct Personalized Query Expansion

PBR Fusion： Integrates stylistic pseudo-feedback with the query and user anchor.

- **Personalized and query-specific dynamic weight**

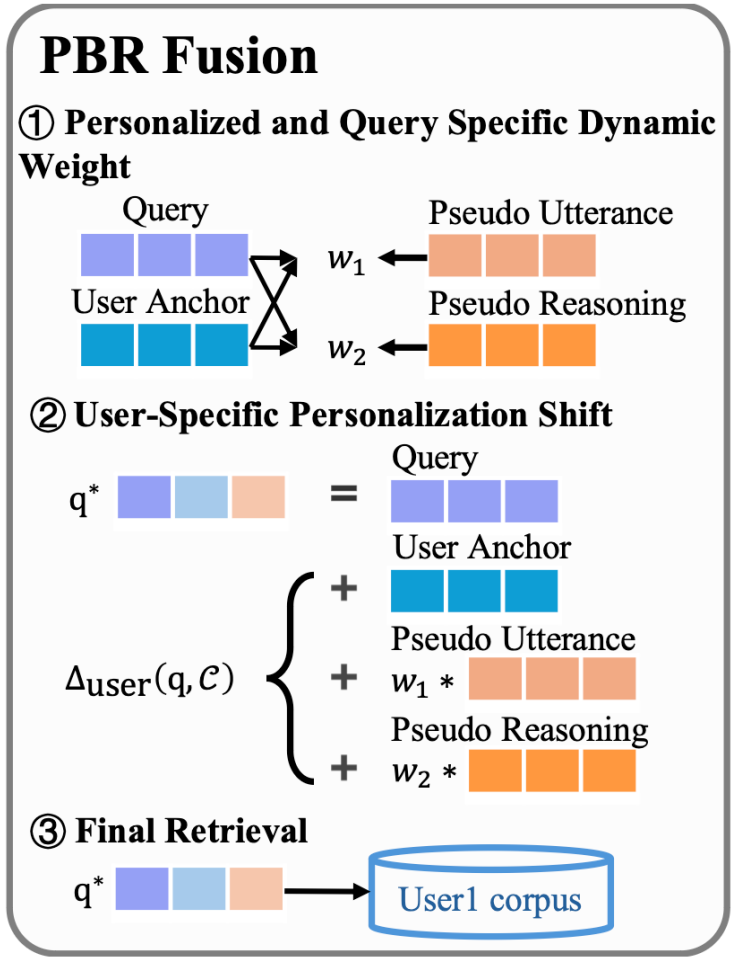
$$w_1 = 1 + \mathrm{sim}\left(\frac{\mathbf{q} + \mathbf{c}_{\mathrm{Anchor}}}{2}, \mathbf{f}_{\mathrm{avg}}\right), \quad w_2 = 1 + \mathrm{sim}\left(\frac{\mathbf{q} + \mathbf{c}_{\mathrm{Anchor}}}{2}, \mathbf{r}\right),$$

- **User-specific personalization shift**

$$\Delta_{\mathrm{user}}(\mathbf{q}, \mathcal{C}) = \mathbf{c}_{\mathrm{Anchor}} + w_1 \cdot \mathbf{f}_{\mathrm{avg}} + w_2 \cdot \mathbf{r},$$

- **Final retrieval** in FAISS-based search index

$$\mathbf{q}^* = \mathbf{q} + \Delta_{\mathrm{user}}(\mathbf{q}, \mathcal{C})$$



Direct Personalized Query Expansion

Case Study

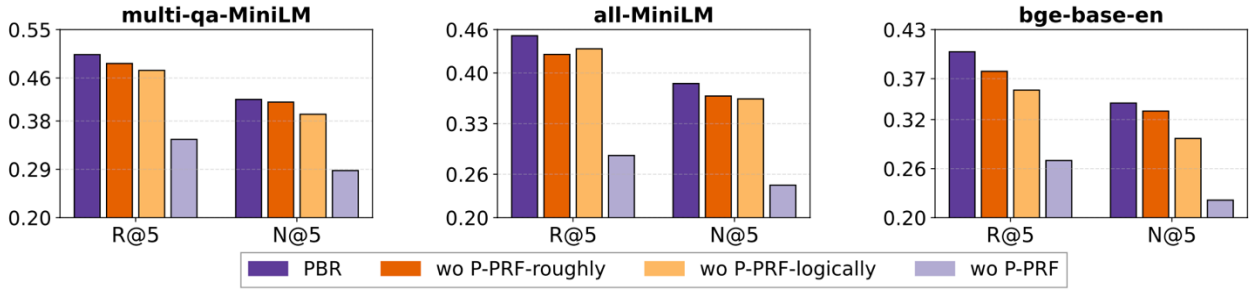
- **P-PRF drives major performance gains in personalized retrieval.** This underscores the importance of P-PRF in generating stylistically aligned pseudo information

Metrics	PBR		w/o P-Anchor		w/o P-PRF		w/o both	
	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5
multi-qa-MiniLM-L6-cos-v1								
Overall	**0.5035**	**0.4201**	0.4871	0.4100	0.3457	0.2877	0.4484	0.3669
Basic information	**0.5091**	**0.3647**	0.4833	0.3570	0.3318	0.2242	0.4515	0.3088
Preference (hard)	0.4049	0.4175	**0.4390**	**0.4295**	0.2683	0.2981	0.3659	0.3759
Social	**0.5541**	**0.4914**	0.5164	0.4529	0.4277	0.3851	0.4852	0.4356
Preference (easy)	**0.5321**	**0.5129**	0.5192	0.5160	0.3590	0.3414	0.4904	0.4582
all-MiniLM-L6-v2								
Overall	**0.4516**	**0.3855**	0.4382	0.3729	0.2860	0.2449	0.3783	0.3074
Basic information	**0.4515**	**0.3485**	0.4379	0.3393	0.1985	0.1617	0.3515	0.2644
Preference (hard)	**0.4341**	**0.4352**	0.4244	0.4201	0.2878	0.2718	0.4000	0.3547
Social	**0.4494**	**0.3777**	0.4333	0.3639	0.4179	0.3387	0.3921	0.3048
Preference (easy)	**0.4840**	**0.4800**	0.4712	0.4587	0.3846	0.3634	0.4295	0.4199
BAAI/bge-base-en-v1.5								
Overall	**0.4029**	**0.3402**	0.3921	0.3256	0.2699	0.2214	0.3738	0.3015
Basic information	**0.4121**	**0.3057**	0.3970	0.2958	0.2576	0.1804	0.3970	0.2748
Preference (hard)	0.3707	**0.3889**	**0.3902**	0.3778	0.2146	0.2173	0.3268	0.3343
Social	**0.3657**	**0.3089**	0.3431	0.2822	0.3261	0.2816	0.3204	0.2799
Preference (easy)	**0.4904**	**0.4734**	0.4744	0.4582	0.2949	0.2785	0.4583	0.4065
Average	**0.4527**	**0.3819**	0.4391	0.3695	0.3005	0.2513	0.4002	0.3253

Direct Personalized Query Expansion

Case Study

- **P-PRF drives major performance gains in personalized retrieval.** This underscores the importance of P-PRF in generating stylistically aligned pseudo information
- **Both Pseudo Utterance and Pseudo Reason are effective.** The Utterance harms lexical coverage, while the Reason impairs goal-oriented reasoning



Direct Personalized Query Expansion

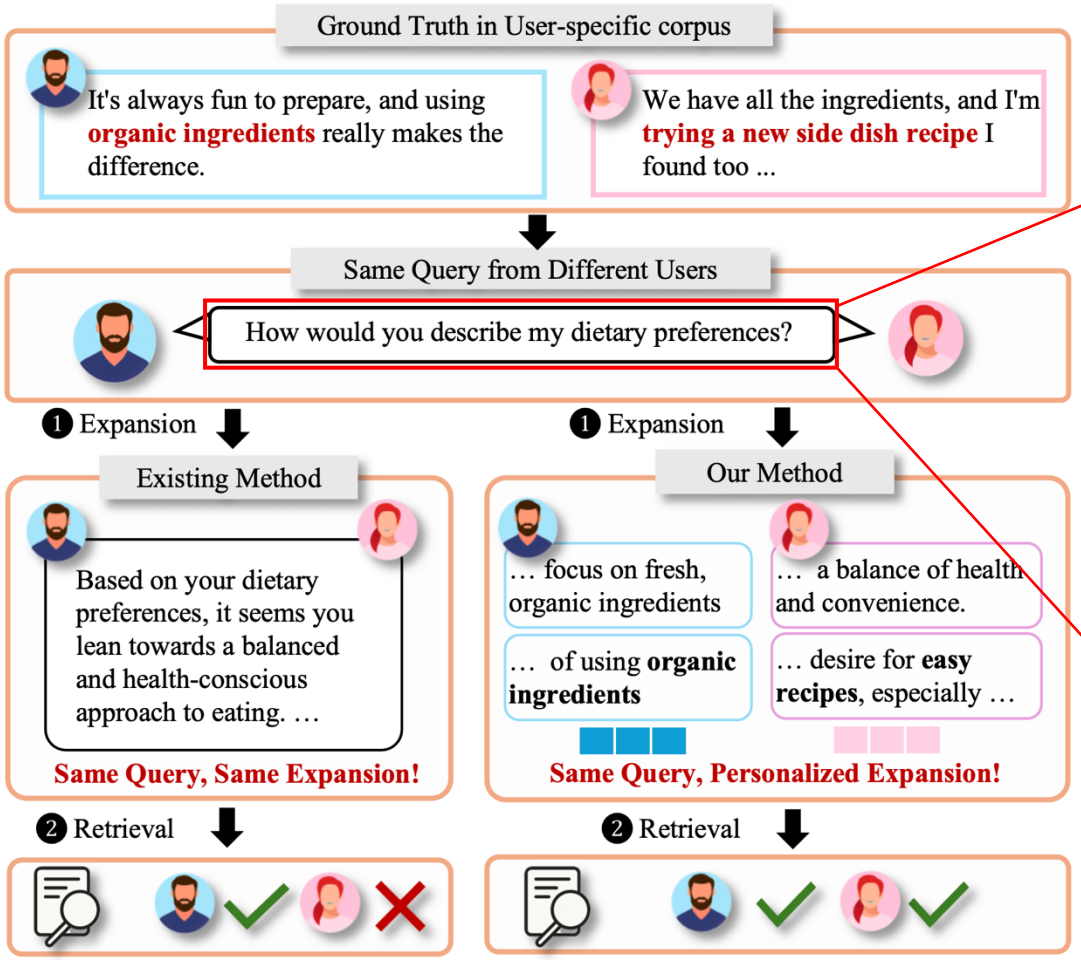
Case Study

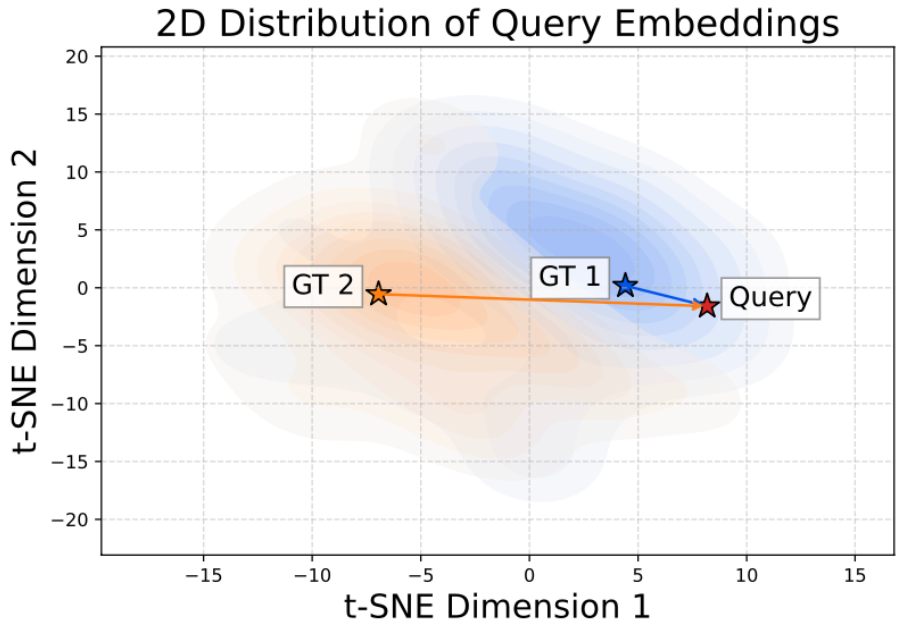
- **P-PRF drives major performance gains in personalized retrieval.** This underscores the importance of P-PRF in generating stylistically aligned pseudo information
- **Both Pseudo Utterance and Pseudo Reason are effective.** The Utterance harms lexical coverage, while the Reason impairs goal-oriented reasoning
- **P-Anchor promotes semantic alignment in structured corpora**. Removing P-Anchor leads to a consistent drop, especially in Basic Information and Social, where user corpora exhibit clearer structure

Metrics	PBR		w/o P-Anchor		w/o P-PRF		w/o both	
	R@5	N@5	R@5	N@5	R@5	N@5	R@5	N@5
multi-qa-MiniLM-L6-cos-v1								
Overall	**0.5035**	**0.4201**	0.4871	0.4100	0.3457	0.2877	0.4484	0.3669
Basic information	**0.5091**	**0.3647**	0.4833	0.3570	0.3318	0.2242	0.4515	0.3088
Preference (hard)	0.4049	0.4175	**0.4390**	**0.4295**	0.2683	0.2981	0.3659	0.3759
Social	**0.5541**	**0.4914**	0.5164	0.4529	0.4277	0.3851	0.4852	0.4356
Preference (easy)	**0.5321**	**0.5129**	0.5192	0.5160	0.3590	0.3414	0.4904	0.4582
all-MiniLM-L6-v2								
Overall	**0.4516**	**0.3855**	0.4382	0.3729	0.2860	0.2449	0.3783	0.3074
Basic information	**0.4515**	**0.3485**	0.4379	0.3393	0.1985	0.1617	0.3515	0.2644
Preference (hard)	**0.4341**	**0.4352**	0.4244	0.4201	0.2878	0.2718	0.4000	0.3547
Social	**0.4494**	**0.3777**	0.4333	0.3639	0.4179	0.3387	0.3921	0.3048
Preference (easy)	**0.4840**	**0.4800**	0.4712	0.4587	0.3846	0.3634	0.4295	0.4199
BAAI/bge-base-en-v1.5								
Overall	**0.4029**	**0.3402**	0.3921	0.3256	0.2699	0.2214	0.3738	0.3015
Basic information	**0.4121**	**0.3057**	0.3970	0.2958	0.2576	0.1804	0.3970	0.2748
Preference (hard)	0.3707	**0.3889**	**0.3902**	0.3778	0.2146	0.2173	0.3268	0.3343
Social	**0.3657**	**0.3089**	0.3431	0.2822	0.3261	0.2816	0.3204	0.2799
Preference (easy)	**0.4904**	**0.4734**	0.4744	0.4582	0.2949	0.2785	0.4583	0.4065
Average	**0.4527**	**0.3819**	0.4391	0.3695	0.3005	0.2513	0.4002	0.3253

Direct Personalized Query Expansion

➢Case Study



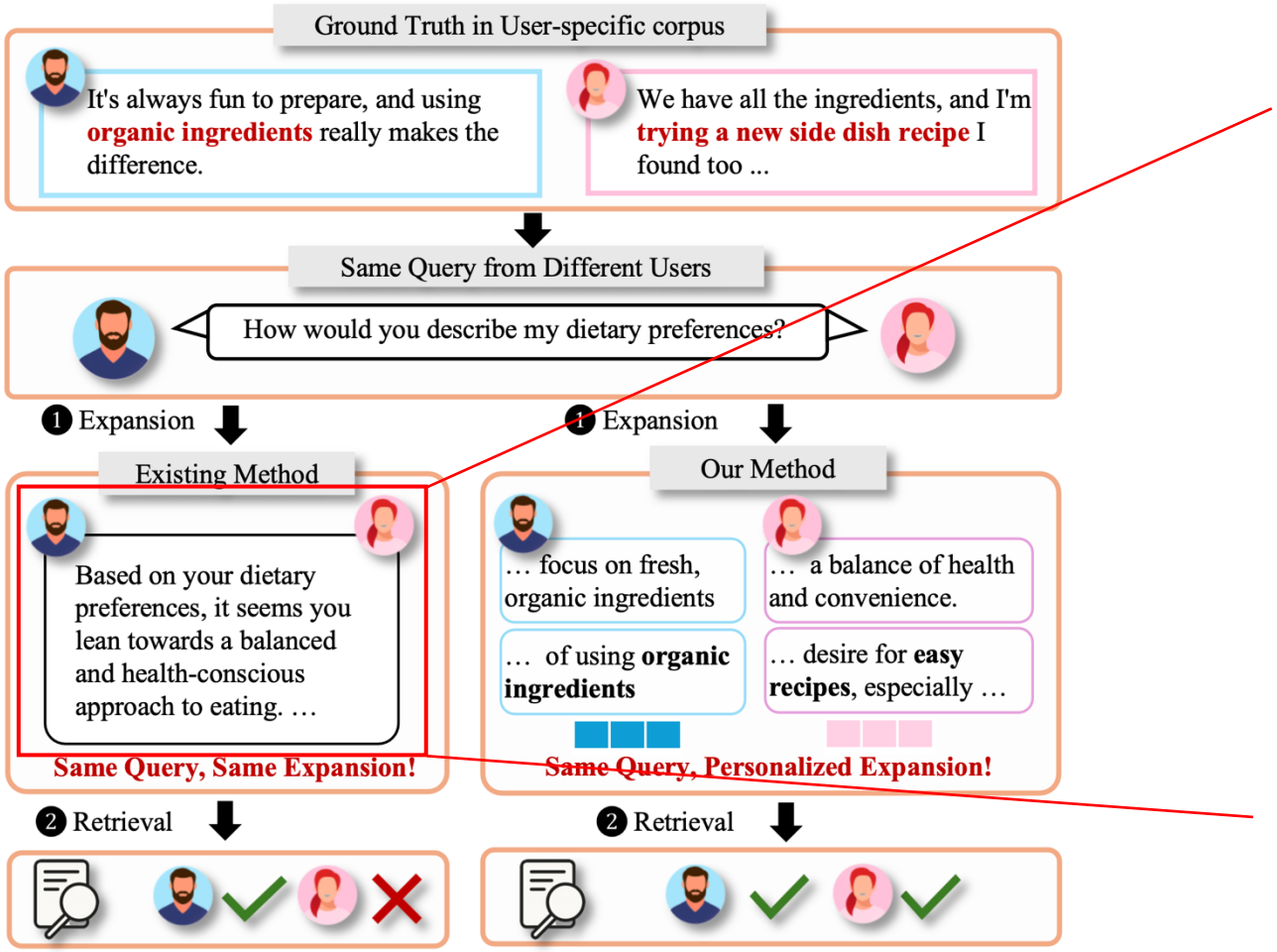


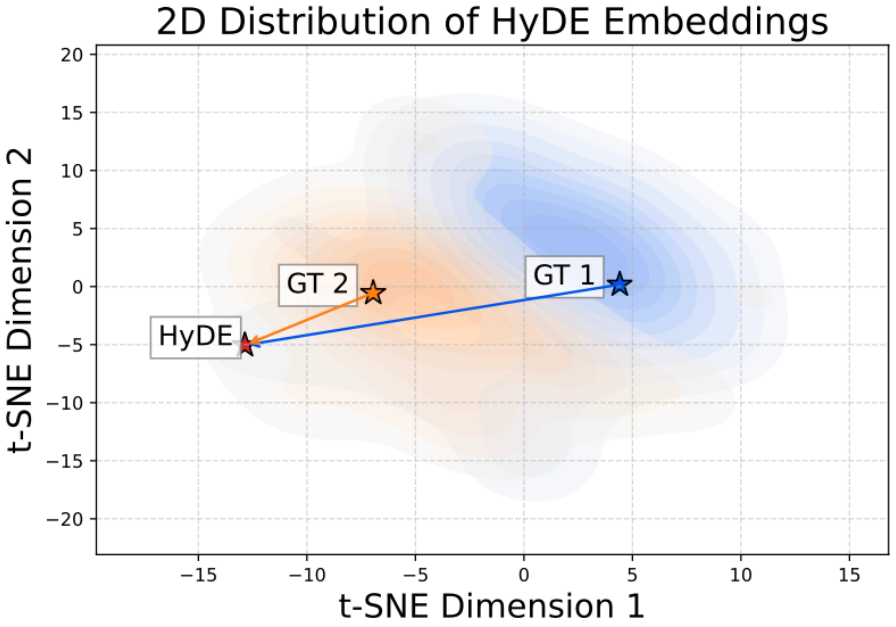
(a) Query-GT Comparison

Using Query to retrieval

Direct Personalized Query Expansion

➢Case Study



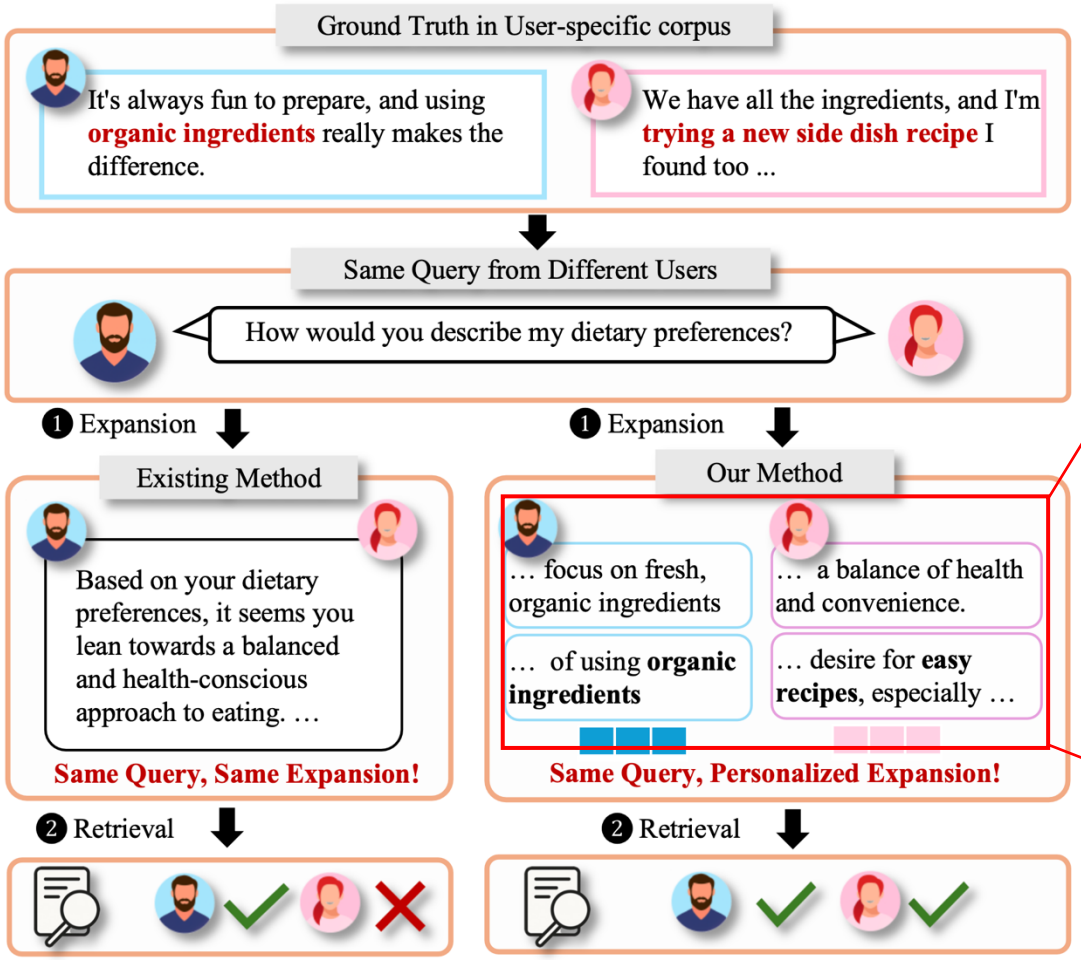


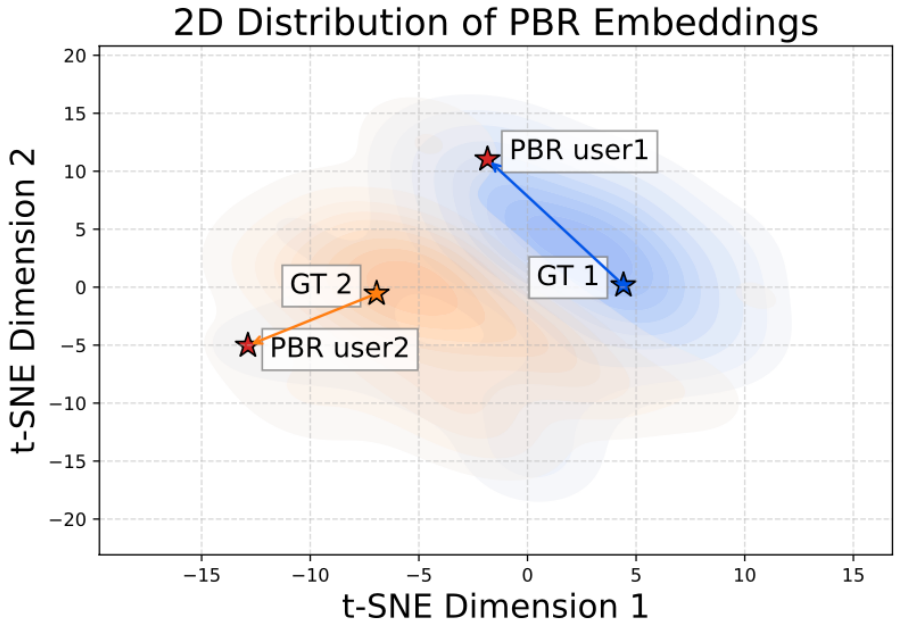
(b) HyDE-GT Comparison

Using HyDE to retrieval

Direct Personalized Query Expansion

➢Case Study





(c) PBR-GT Comparison (Ours)

Use PBR Retrieval

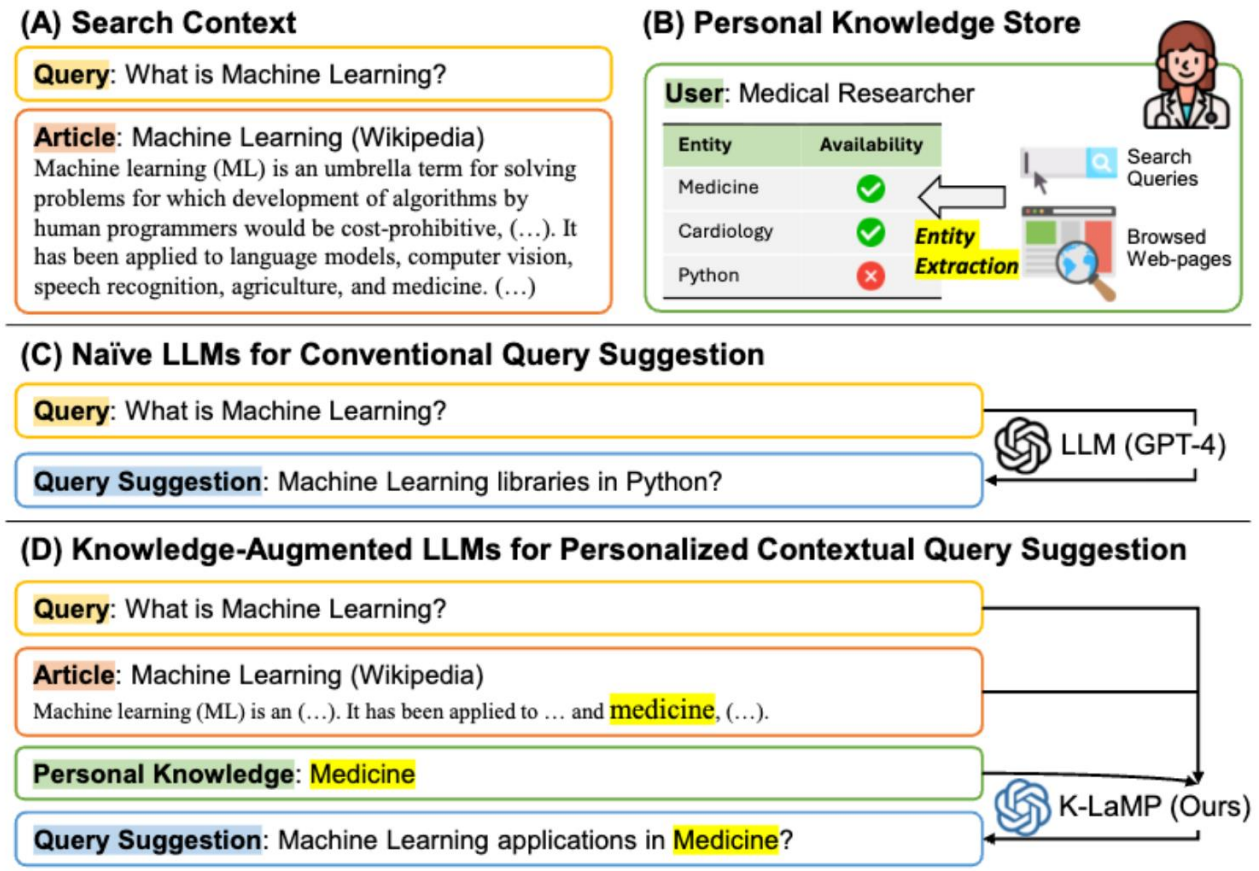
Others

- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query

- K-LaMP (WWW'24)

  - **Problem:**
    - LLMs are powerful but static, need current context and cannot capture user-specific knowledge

  - **QE method Require**
    - Personalization is need via knowledge augmentation, not model retraining

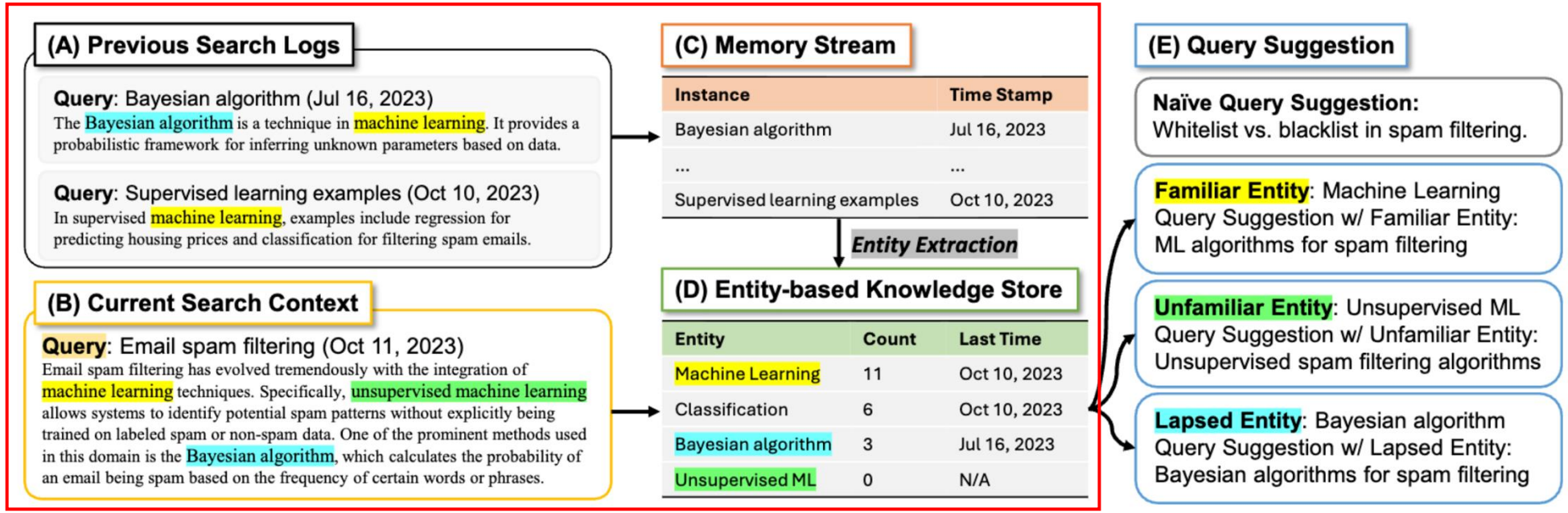


[1] Baek, Jinheon, et al. "Knowledge-augmented large language models for personalized contextual query suggestion." *Proceedings of the ACM Web Conference 2024*

Others

- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query

- K-LaMP (WWW'24)
  - **1. Build Personal Knowledge Store**
    - Extract entities from user queries, browsing history and map them to a public knowledge graph

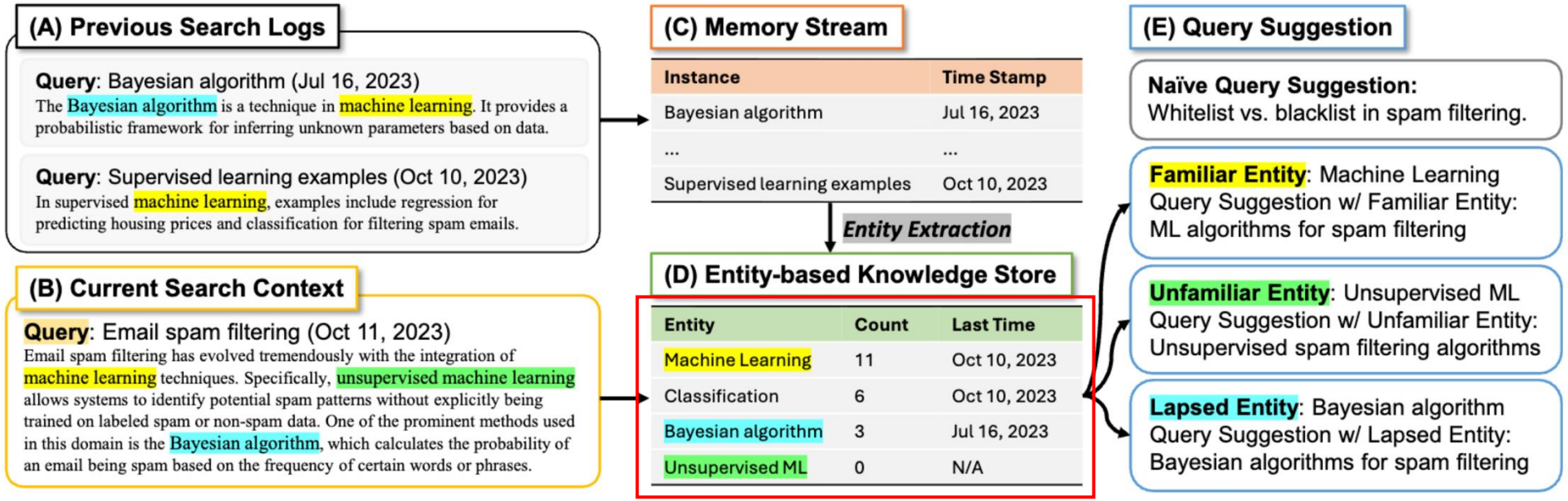


[1] Baek, Jinheon, et al. "Knowledge-augmented large language models for personalized contextual query suggestion." *Proceedings of the ACM Web Conference 2024*

Others

- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query

- K-LaMP (WWW'24)
  - **2. Entity-Centric Representation**
    - Store user knowledge as aggregated entity projections to avoid storing raw logs and improves scalability

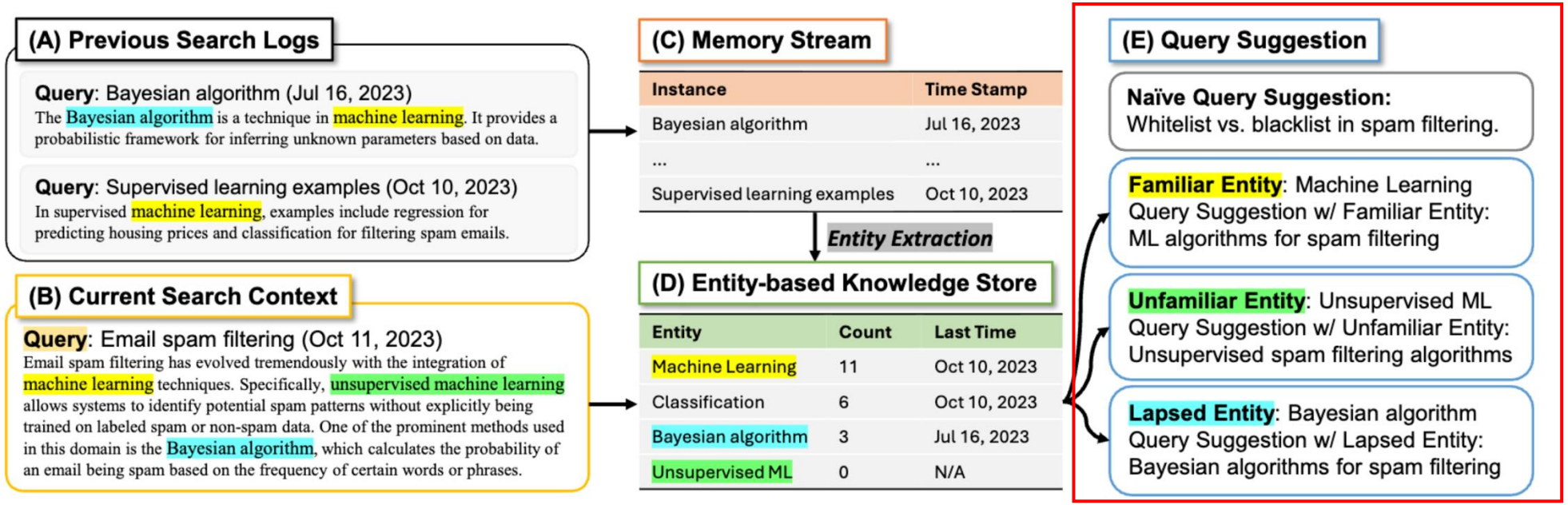


[1] Baek, Jinheon, et al. "Knowledge-augmented large language models for personalized contextual query suggestion." *Proceedings of the ACM Web Conference 2024*

Others

- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query

- K-LaMP (WWW'24)
  - **3. Prompt Augmentation and Query Suggestion Generation**
    - Retrieve relevant entities from knowledge store and Inject into LLM prompt as context
    - Generate contextually relevant queries



[1] Baek, Jinheon, et al. "Knowledge-augmented large language models for personalized contextual query suggestion." *Proceedings of the ACM Web Conference 2024*

Others

- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query

- K-LaMP (WWW'24)
  - **Commercial and non-contextual methods mainly generate generic suggestions,** such as "Tim Cook Boyfriend," "Tim Cook Latest News," or "Tim Cook net worth."
  - **K-LaMP generates more personalized and context-aware suggestions,** such as Tim Cook's impact on Apple's product line and the difference between Tim Cook's and Steve Jobs' approaches.

Table 6: A manufactured example showing the type of data that was provided to human judges.

Types	Texts
Query	Tim Cook
Session	Apple | Tim Cook
Article	Tim Cook Leadership: A new profile examines how Apple CEO Tim Cook, with "cautious, collaborative and tactical" leadership, honed the Cupertino tech giant into the world's largest company. (…)
Trending entities	'GPT-4', 'OpenAI', 'Google Bard', 'Microsoft Copilot', 'Elon Musk', …
Personal summary	The user is interested in Apple products and technology ('Macbook', 'macOS', and 'Apple TV'). They have a keen interest in ML, with topics like 'Supervised Learning' and 'Optimization'. Additionally, they enjoy animation, showing interest in 'Studio Ghibli', 'Walt Disney', and 'Pixar'. Their preferences also extend to home entertainment ('DVD' and 'HDTV'). Lastly, they follow baseball ('MLB' and 'New York Yankees').
Personal entities	'Macbook', 'macOS', 'Machine Learning', 'Optimization', 'Supervised Learning', 'Apple TV', 'Animation', 'Studio Ghibli', 'DVD', 'Walt Disney', 'Pixar Animation Studios', 'Apple Inc.', 'Baseball', 'HDTV', 'Major League Baseball', 'New York Yankees', …

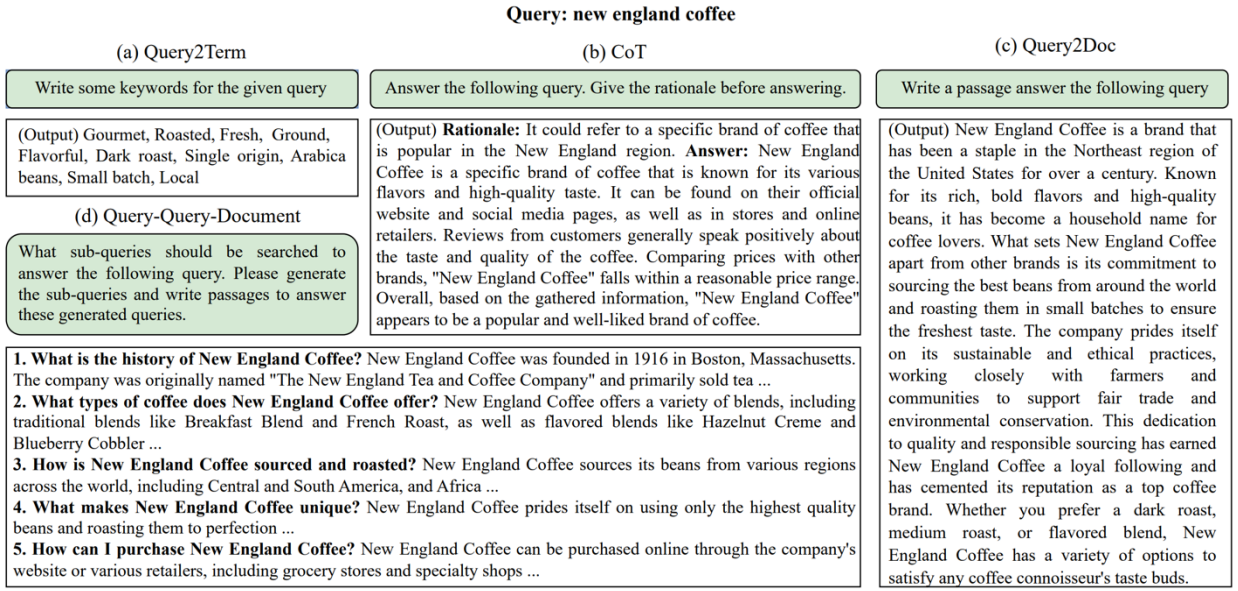
Table 7: Query suggestion results of different approaches including non-contextual and commercial methods, based on the data example in Table 6.

Models	Query Suggestion Results
Microsoft Bing Search	"Tim Cook Boyfriend", "Tim Cook Latest News"
Google Search	"Tim Cook Quotes", "Tim Cook Boyfriend"
Query Suggestion w/o Context	"Tim Cook leadership style", "Tim Cook net worth"
K-LaMP (Ours)	"Tim Cook's impact on Apple's product line", "Difference between Tim Cook's and Steve Jobs' approach to Apple's product development"

[1] Baek, Jinheon, et al. "Knowledge-augmented large language models for personalized contextual query suggestion." *Proceedings of the ACM Web Conference 2024*

Others

- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query
- MILL (NAACL'24 Our Method)
  - **Problem:**
    - Existing QE Depends on top retrieved documents
    - Fails when query is short or initial retrieval is poor and lacks corpus grounding or may hallucinate
  - **QE method Require**
    - Combine generation and retrieval via mutual verification



[1] Jia, Pengyue, et al. "Mill: Mutual verification with large language models for zero-shot query expansion." *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024.

Others

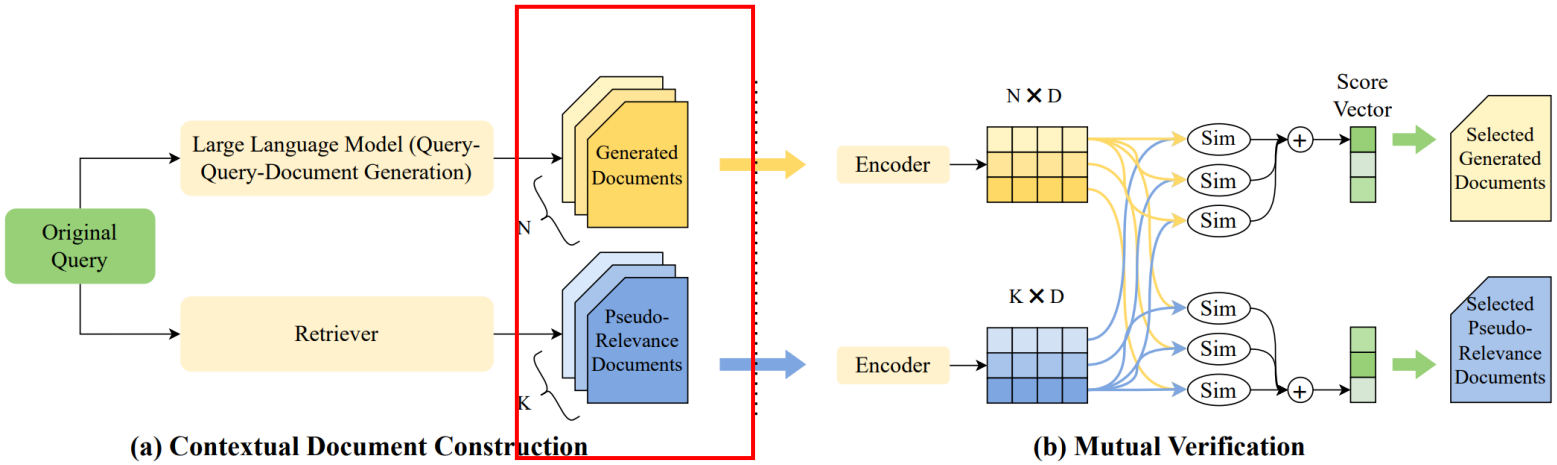
- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query

- MILL (NAACL'24 Our Method)

  - **1. Query–Query–Document Generation (QQD)**
    - Use LLMs to decompose the original query into multiple sub-queries, then generate pseudo documents.

What sub-queries should be searched to answer the following query. Please generate the sub-queries and write passages to answer these generated queries.

$$\mathcal{D}^{\text{LLM}} = \{d_n^{\text{LLM}}\} = \text{LLM}(p, q),\ n \in (0, N]$$
$$\mathcal{D}^{\text{PRF}} = \{d_k^{\text{PRF}}\} = R_r(q),\ k \in (0, K]$$



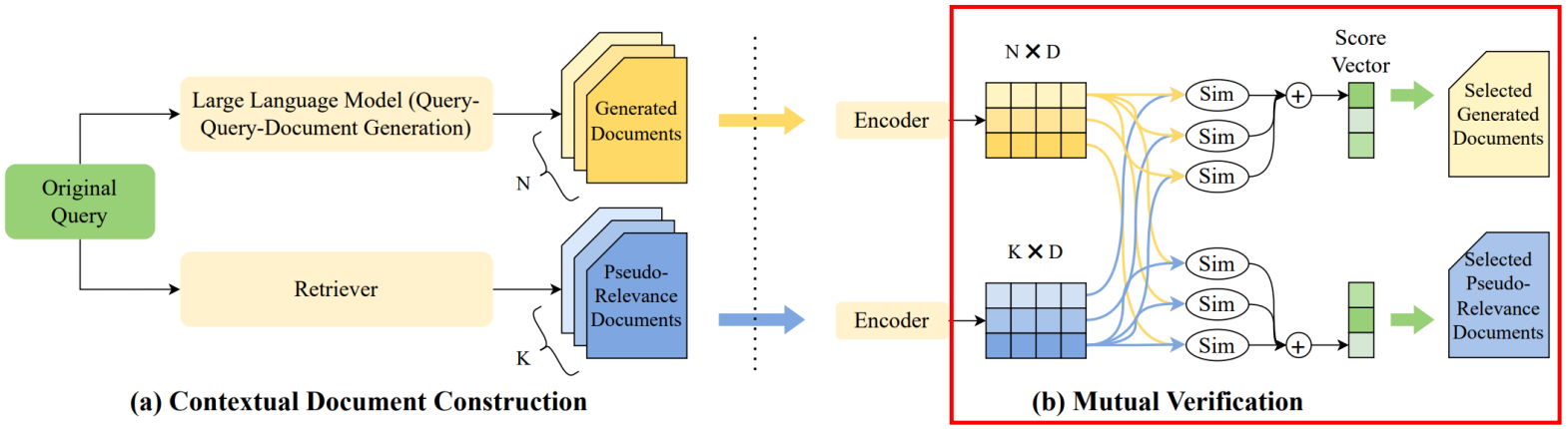
[1] Jia, Pengyue, et al. "Mill: Mutual verification with large language models for zero-shot query expansion." *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024.

Others

- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query
- MILL (NAACL'24 Our Method)
  - **2. Mutual Verification**
    - Instead of trusting generated or retrieved content alone
    - Let Generated documentsand Retrieved documents verify each other

What sub-queries should be searched to answer the following query. Please generate the sub-queries and write passages to answer these generated queries.

$$\mathcal{D}^{\text{LLM}} = \{d_n^{\text{LLM}}\} = \text{LLM}(p, q),\ n \in (0, N]$$
$$\mathcal{D}^{\text{PRF}} = \{d_k^{\text{PRF}}\} = R_r(q),\ k \in (0, K]$$
$$s_n^{\text{LLM}} = \sum_{k=1}^{K} \text{sim}(\mathbf{x}_n^{\text{LLM}}, \mathbf{x}_k^{\text{PRF}}),$$
$$s_k^{\text{PRF}} = \sum_{n=1}^{N} \text{sim}(\mathbf{x}_k^{\text{PRF}}, \mathbf{x}_n^{\text{LLM}}).$$



[1] Jia, Pengyue, et al. "Mill: Mutual verification with large language models for zero-shot query expansion." *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024.

Others

- Incorporating LLMs' world knowledge, search history, social networks, or preferences derived from friendship networks to expand original query

- MILL (NAACL'24 Our Method)
  - **The LLM-generated expansion content can provide more relevant domain knowledge than the top pseudo-relevance document**, which mistakenly focuses on "nonconformance" rather than the earth science concept "nonconformity."
  - **By filtering out noisy pseudo-relevance documents and less useful LLM-generated content,** the method keeps expansion terms that better match the ground-truth answer, such as erosion, sedimentary rocks, igneous or metamorphic rocks, and geologic record.

Query	What is a nonconformity earth science
Expansion Content	Pseudo-relevance document: There are three types of unconformities(below): angular **unconformity**, **nonconformity**, and disconformity. Anangular **unconformity** (left) is an erosional **surface** separating steeply dipping **rock** layers below from gently dipping layers above. **Unconformity** is the general term for **missing** pages of Earth history. LLM-generated document: **1.** What is the definition of a **nonconformity** in earth science? A **nonconformity** in earth science refers to a type of **unconformity**, which is a **break in the geologic record** where layers of **rock** are **missing**. Specifically, a **nonconformity** is where **sedimentary rocks** are deposited on top of **igneous or metamorphic rocks**. **2.** How are **nonconformities** formed inearth science? **Nonconformities** are formed throughthe process of **erosion**, where layers of **rock** are gradually wornaway. This exposes the underlying **igneous or metamorphic rocks**, which are then **covered by** new **sedimentary rocks** over time. **3.** What are the different types of **nonconformities** in earth science? There are three main types of **nonconformities**: angular, **nonconformity**, and disconformity. Angular **nonconformities occur** when **sedimentary rocks** are deposited at an angle on top of tilted or folded **igneous or metamorphic rocks**. **Nonconformities occur** when **sedimentary rocks** are deposited on top of **eroded igneous ormetamorphic rocks**. Disconformities **occur** when there is a gap in the **sedimentary rock** layers, indicating **a period of erosion** or non-deposition. **4.** What are some examples of **nonconformities found** in the Earth's geology? **One well-known example** of a **nonconformity** is the **Grand Canyon** in Arizona, USA. The **sedimentary layers** seen in the **canyon** were deposited on top of tilted and folded igneous and **metamorphic rocks**, indicating a long history of **erosion and deposition**.
Ground Truth	**Nonconformities** are **unconformities** that separate **igneous or metamorphic rocks** from overlying **sedimentary rocks**. They usually indicate that along **period of erosion occurred** prior to **deposition of the sediments** (several km of **erosion** necessary). They are a feature of stratified **rocks**, and are therefore usually found in **sediments** (but may also occur in stratified volcanics). They are **surfaces** between two **rock** bodies that constitute a substantial **break (hiatus) in the geologic record** (sometimes people say inaccurately that time is **missing**). **Nonconformity**. When **igneous or metamorphic rocks** are **eroded** and then **covered by** younger **sedimentary rocks**, the contact is called a **nonconformity**. **One of the most famous** of these is **found in the Grand Canyon**, where the oldest **sedimentary rocks** are more than a billion years younger than the 1.6 billion-year-old **metamorphic rocks** on which they rest.
Filtered-out PRF document (ranked #1 by BM25 with the original query)	Definition of nonconformance in the AudioEnglish.org Dictionary. Meaning of non-conformance. What does nonconformance mean? Proper usage of the word nonconformance. Information about nonconformance in the AudioEnglish.org dictionary, synonyms, and antonyms.
Filtered-out LLM-generated document	... **7.** What other geological features are commonly associated with nonconformities in earth science? Nonconformities are often found alongside other geological features, such as faults, folds, and intrusions, which can all provide additional information about the Earth's history and the processes that have shaped it. **8.** How can nonconformities in earth science be identified in the field? Nonconformities can be identified by looking for the distinct contact between two different rock types, as well as the difference in age between the two layers. Geologists may also use specialized tools, such as radiometric dating, to determine the age of the rocks. **9.** Are nonconformities only found on land in earth science? No, nonconformities can also be found underwater in the oceans, where layers of sedimentary rock are exposed and show similar

[1] Jia, Pengyue, et al. "Mill: Mutual verification with large language models for zero-shot query expansion." *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024.

Pre-retrieval: Query Expansion Overview

Method	Category	Paper
Gossple	Tagging	Bertier, Marin, et al. "Toward personalized query expansion." Proceedings of the Second ACM EuroSys Workshop on Social Network Systems. 2009.
Biancalana, Claudio et al.	Tagging	Biancalana, Claudio, and Alessandro Micarelli. "Social tagging in query expansion: A new way for personalized web search." 2009 International Conference on Computational Science and Engineering. Vol. 4. IEEE, 2009.
SoQuES	Tagging	Bouadjenek, Mohamed Reda, et al. "Personalized social query expansion using social bookmarking systems." Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. 2011.
Zhou, Dong et al.	Tagging	Zhou, Dong, Séamus Lawless, and Vincent Wade. "Improving search via personalized query expansion using social media." Information retrieval 15 (2012): 218-242.
Query2Doc	Else	Wang, Liang, Nan Yang, and Furu Wei. "Query2doc: Query expansion with large language models." arXiv preprint arXiv:2303.07678 (2023).
MILL	Else	Jia, Pengyue, et al. "MILL: Mutual Verification with Large Language Models for Zero-Shot Query Expansion." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024.
Lin, Shan-Mu et al.	Else	Lin, Shan-Mu, and Chuen-Min Huang. "Personalized optimal search in local query expansion." Proceedings of the 18th Conference on Computational Linguistics and Speech Processing. 2006.
PQEC	Else	Philippe Mulhem, Nawal Ould Amer, and Mathias Géry. 2016. Axiomatic term-based personalized query expansion using bookmarking system. In International Conference on Database and Expert Systems Applications. Springer, 235–243.
WE-LM	Else	Xuan Wu, Dong Zhou, Yu Xu, and Séamus Lawless. 2017. Personalized query expansion utilizing multi-relational social data. In 2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP). IEEE, 65–70.
PSQE	Else	Mohamed Reda Bouadjenek, Hakim Hacid, and Mokrane Bouzeghoub. 2019. Personalized social query expansion using social annotations. Transactions on Large-Scale Data-and Knowledge-Centered Systems XL (2019), 1–25.
PQEWC	Else	Elias Bassani, Nicola Tonellotto, and Gabriella Pasi. 2023. Personalized query expansion with contextual word embeddings. ACM Transactions on Information Systems 42, 2 (2023), 1–35.

Pre-retrieval Takeaways

- The key takeaways for pre-retrieval are listed as follows
  - **Query rewriting** is most beneficial when the original query is **ambiguous**, underspecified, or misaligned with retrieval intents, particularly in conversational or multi-turn settings.
  - **Query expansion** is most effective when the original query is **relevant but incomplete** — i.e., when it needs to be semantically broadened to cover additional relevant concepts.

Agenda

1. Introduction
9:00 – 9:50



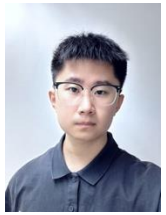
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

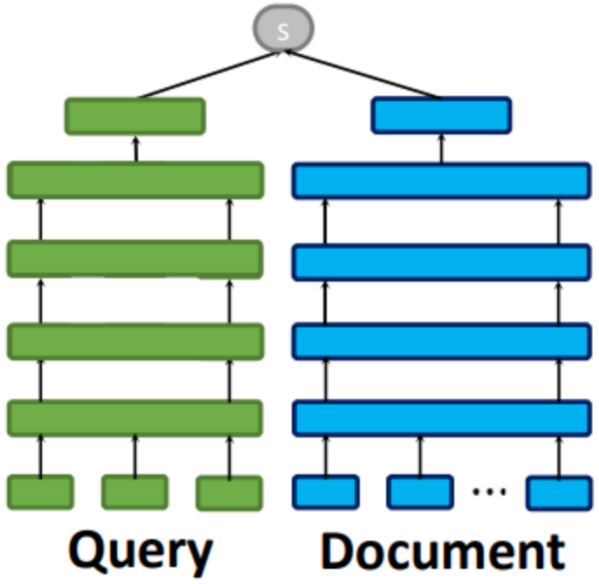
8. Hands on Projects
16:45 – 17:30



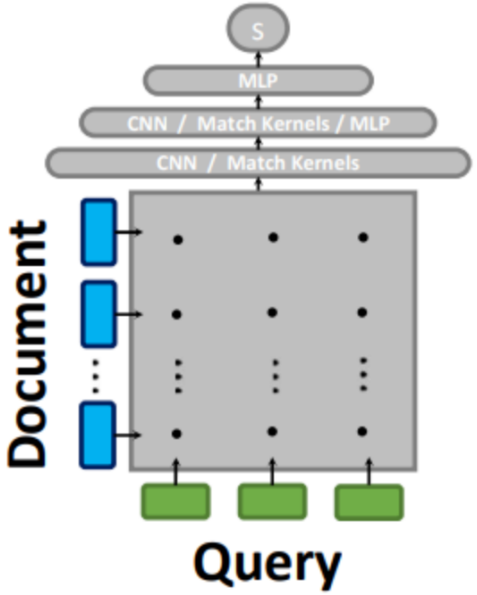
Pengyue Jia

Retrieval: Problem

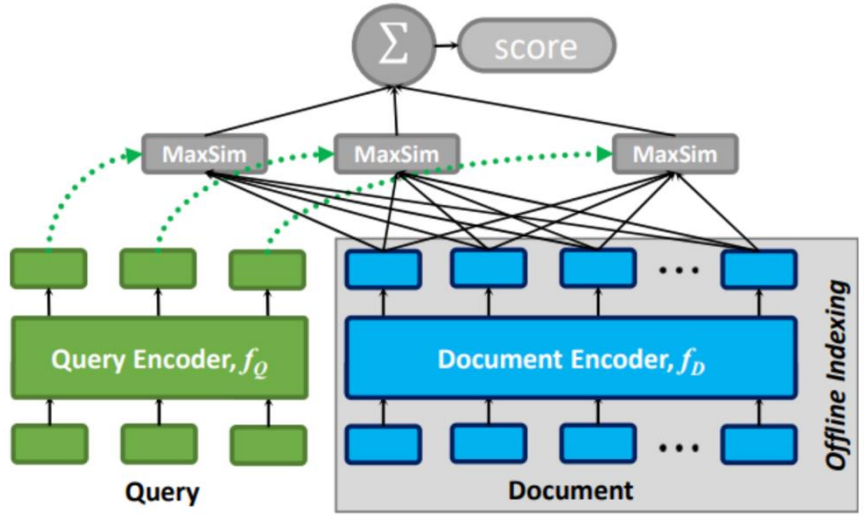
Existing retrieval methods use a shared, agnostic similarity model to match queries and documents.



DSSM(CIKM, Microsoft)



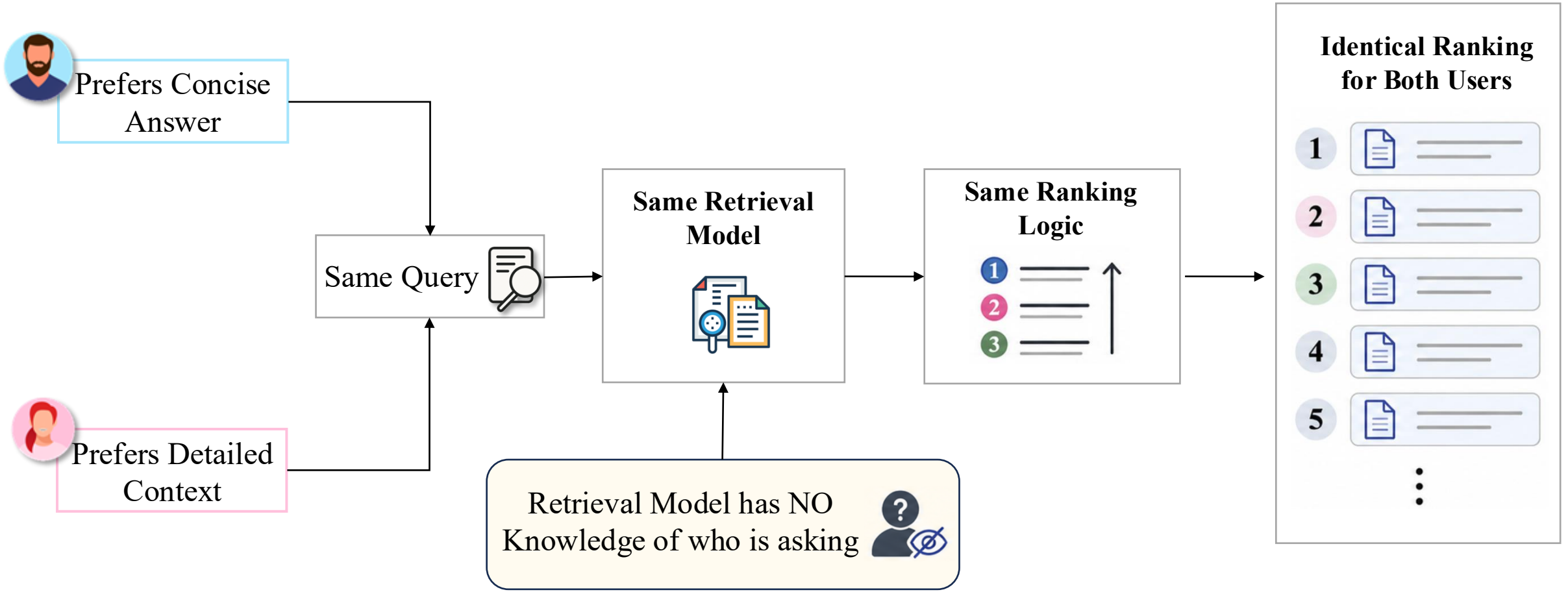
K-NRM(SIGIR, CMU)



ColBERT (SIGIR, Stanford)

Retrieval: Problem

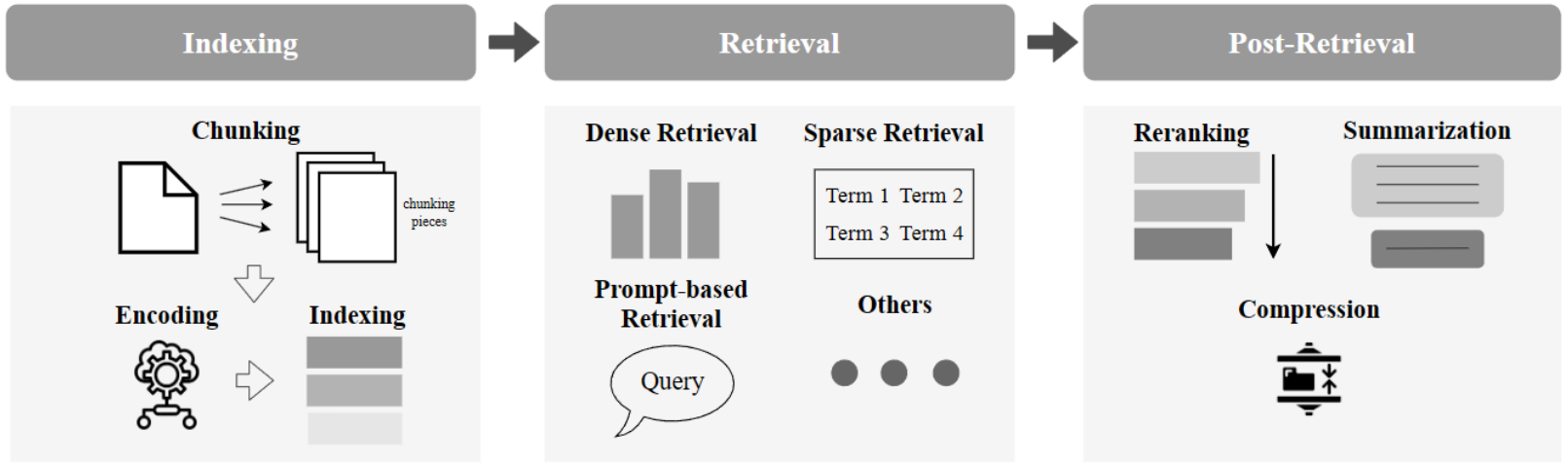
A shared retrieval model scores all users equally



Retrieval: Formula definition

- Retrieval Definition:
  - The retrieval process involves finding the most relevant documents D^* from a corpus C based on a query q^*. Personalization is integrated into the retrieval function $\mathcal{R}$ using user-specific information p, enhancing relevance and user preference alignment.

- Equation: $D^* = \mathcal{R}(q^*, C, p)$
  - p is the personalization information of the user
  - q^* is the enhanced query
  - C is the corpus



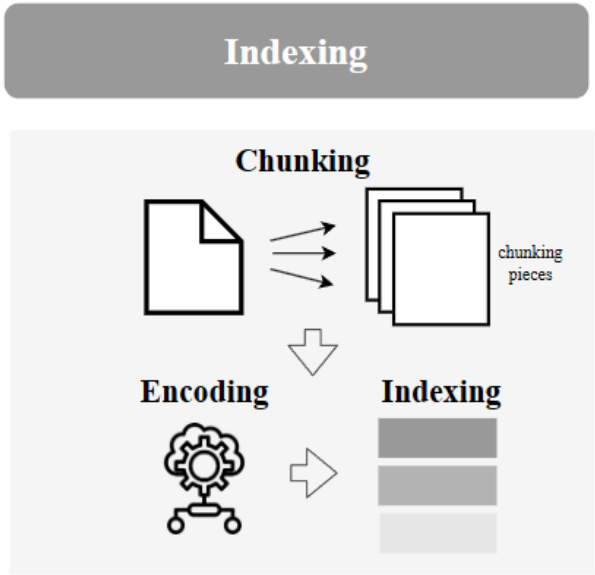
- Retrieval catrgories:
  - Indexing
  - Retrieval
  - Post-Retrieval

Retrieval: Indexing

- Motivation
  - ***Enhance retrieval accuracy by tailoring results to individual user needs.***

- Personalized Indexing Definition
  - Organizing knowledge base data into structured formats while incorporating user-specific embeddings or preferences for efficient and tailored retrieval.

- Methods can be divided to:
  - Chunking
  - Encoding
  - Indexing



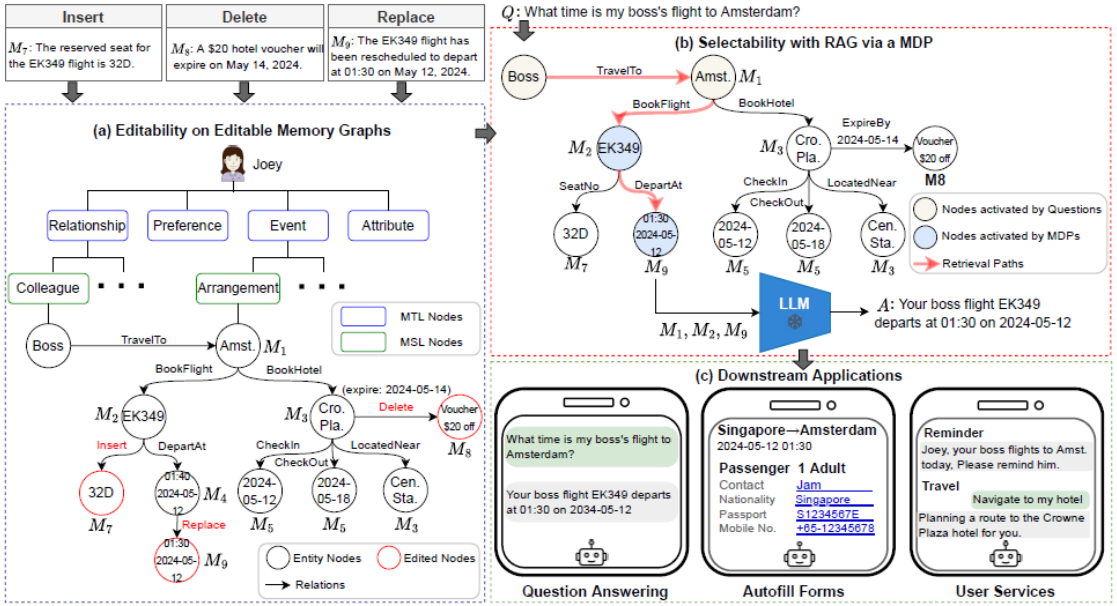
Indexing

➢ EMG-RAG (EMNLP'24)

- ➢ Enables personalized indexing by building a **user-specific, editable memory graph** for each user, replacing the traditional shared static index with a dynamic personal knowledge structure

- ➢ **Problem:**
  - ➢ Personal memories are heterogeneous and hard to organize
  - ➢ User memory is dynamic — static indexes become outdated

- ➢ **Method:**
  - ➢ Constructs a three-layer EMG per user
  - ➢ Supports Insert / Delete / Replace operations directly on the graph index, keeping each user's personal index up-to-date

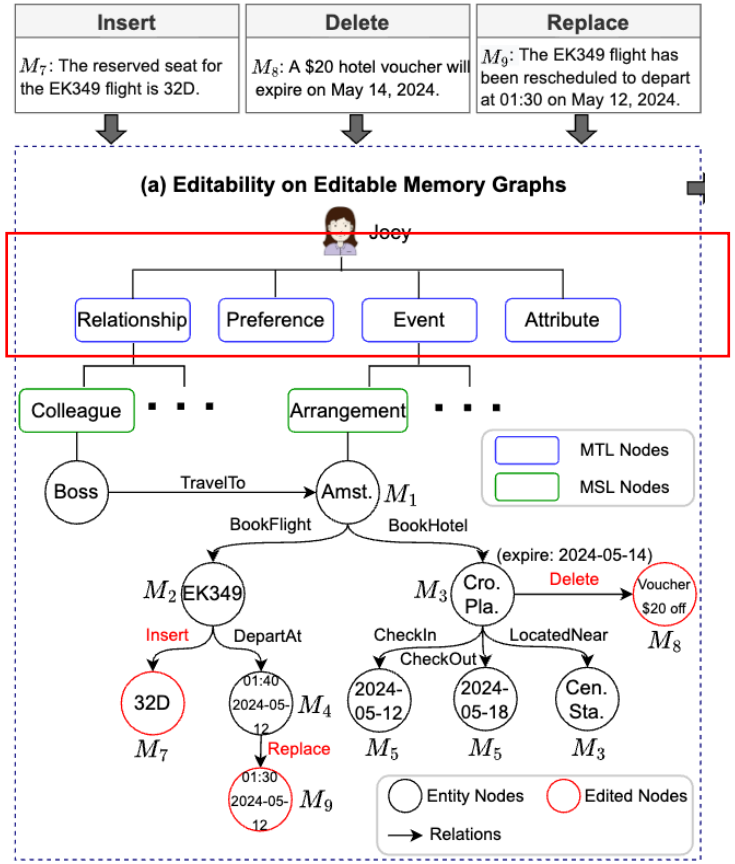


[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

- Focusing on personalized Indexing: Editable Memory Graph

- **EMG-RAG (EMNLP'24)**
  - **1. Three-Layer EMG**
    - Memory Type Layer (MTL): Categorizes memories into 4 types — Relationship, Preference, Event, Attribute.



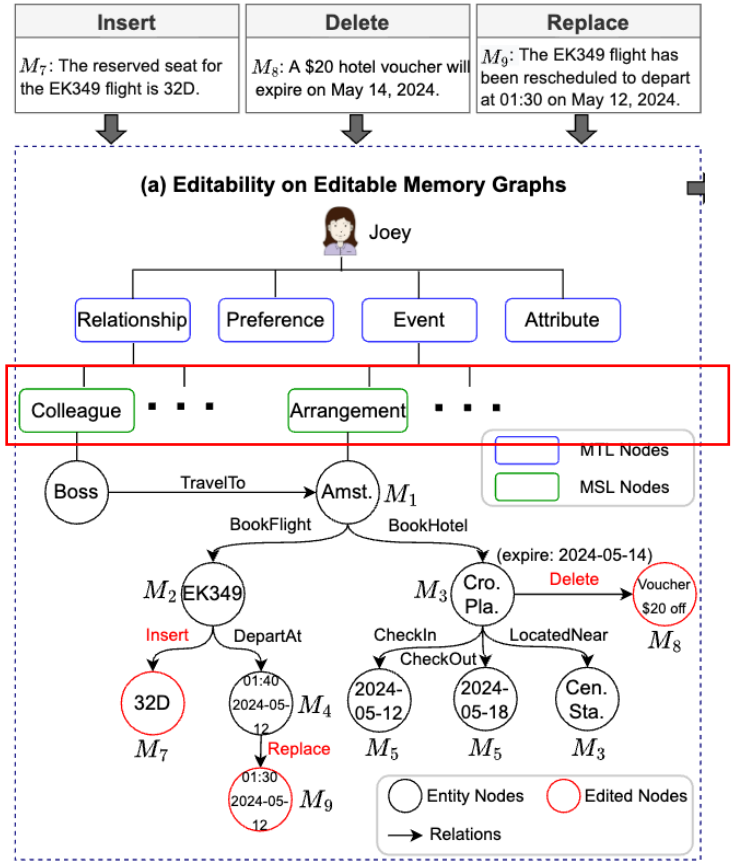
Memory Type Layer (MTL)

[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

- Focusing on personalized Indexing: Editable Memory Graph

- **EMG-RAG (EMNLP'24)**
  - **1. Three-Layer EMG**
    - Memory Type Layer (MTL): Categorizes memories into 4 types — Relationship, Preference, Event, Attribute.
    - Memory Subclass Layer (MSL): Further divides each type into fine-grained subclasses



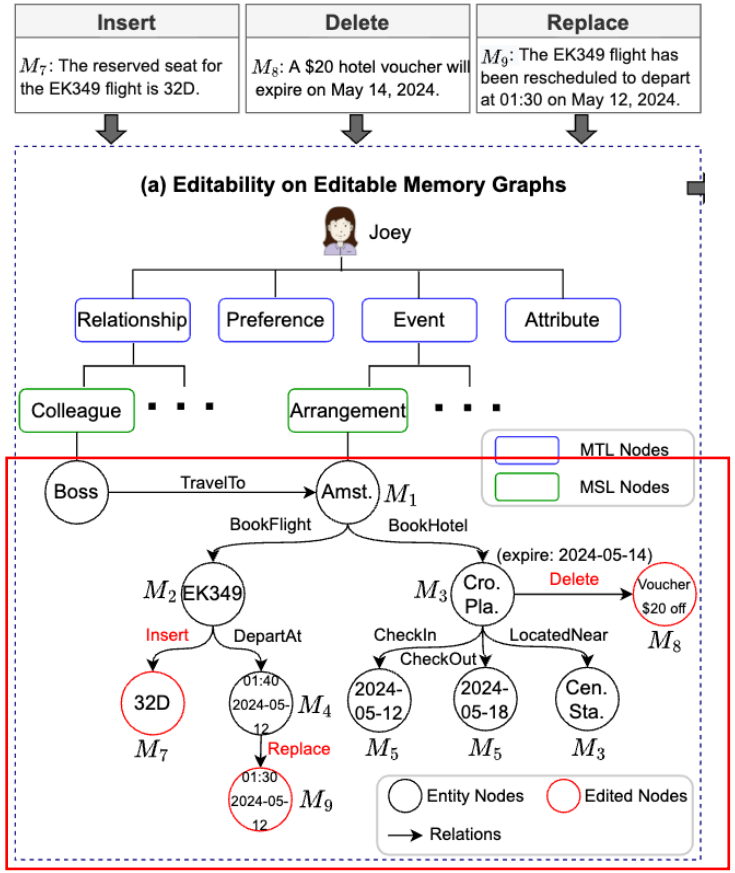
Memory Subclass Layer (MSL)

[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

- Focusing on personalized Indexing: Editable Memory Graph

- **EMG-RAG (EMNLP'24)**
  - **1. Three-Layer EMG**
    - Memory Type Layer (MTL): Categorizes memories into 4 types — Relationship, Preference, Event, Attribute.
    - Memory Subclass Layer (MSL): Further divides each type into fine-grained subclasses
    - Memory Graph Layer (MGL): Stores entity nodes and relational edges parsed from actual memories



Memory Graph Layer (MGL)

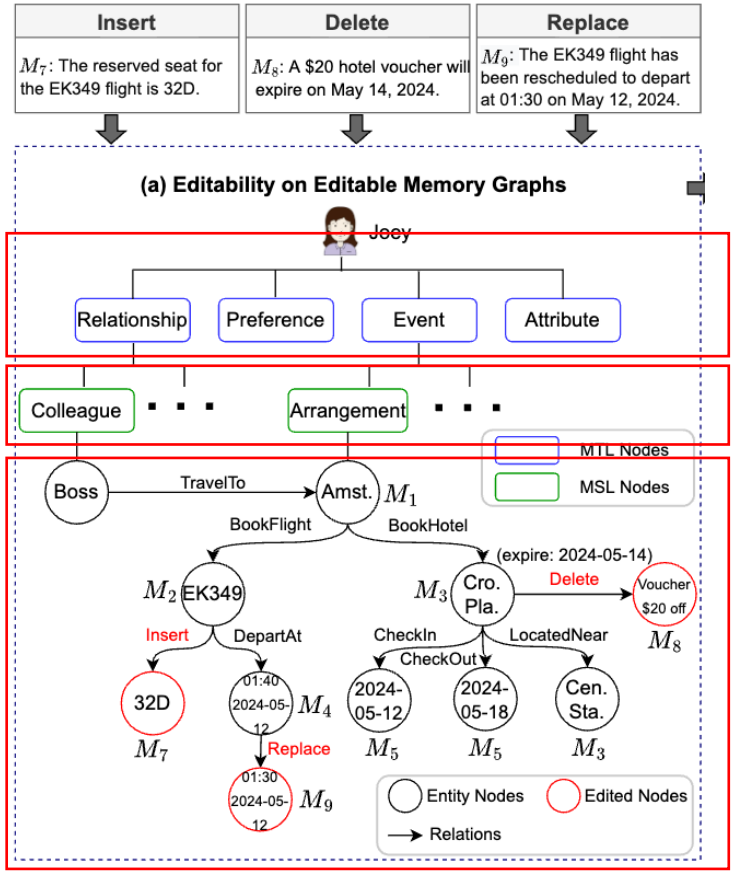
[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

➢ Focusing on personalized Indexing: Editable Memory Graph

➢ EMG-RAG (EMNLP'24)

- ➢ **1. Three-Layer EMG**
  - ➢ Memory Type Layer (MTL): Categorizes memories into 4 types — Relationship, Preference, Event, Attribute.
  - ➢ Memory Subclass Layer (MSL): Further divides each type into fine-grained subclasses
  - ➢ Memory Graph Layer (MGL): Stores entity nodes and relational edges parsed from actual memories
  - ➢ **User independent constructed EMG**



Memory Type Layer (MTL)

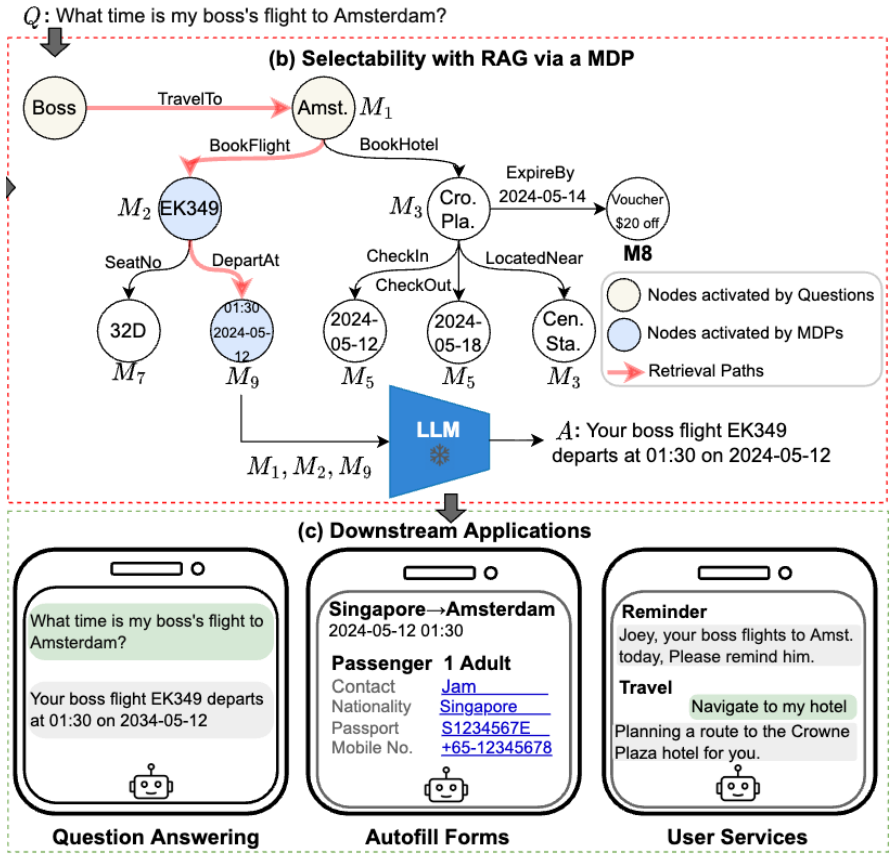
Memory Subclass Layer (MSL)

Memory Graph Layer (MGL)

[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

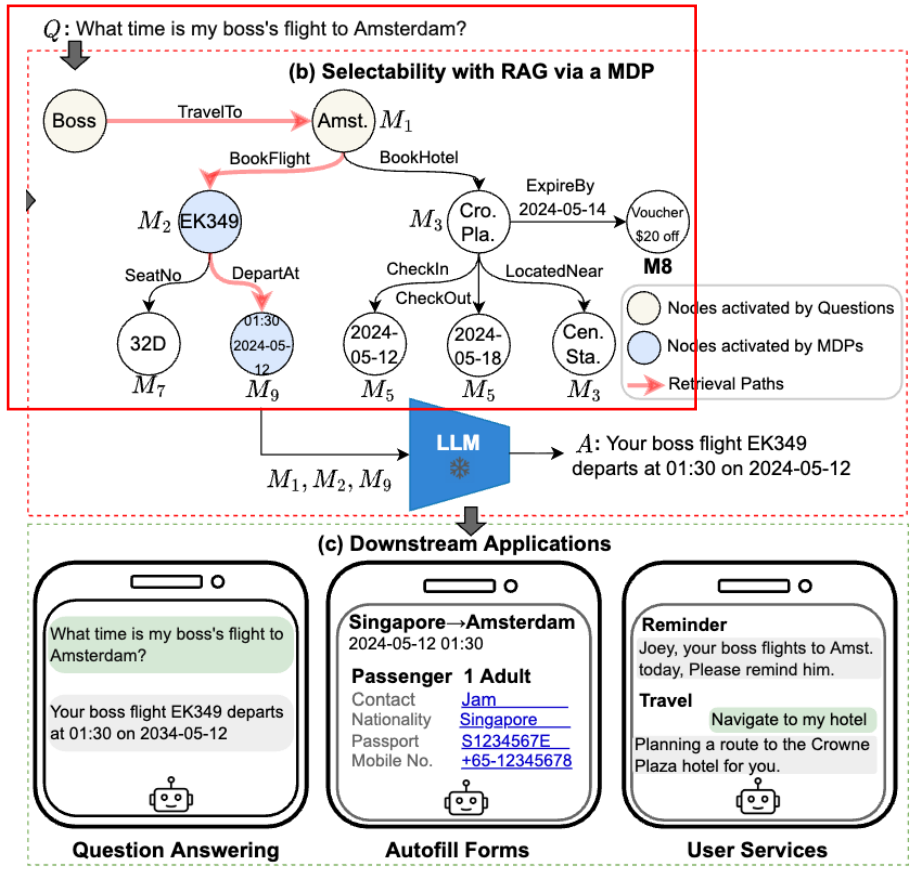
- Focusing on personalized selectability — adaptively retrieving the right combination of memories for each user's query.



[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

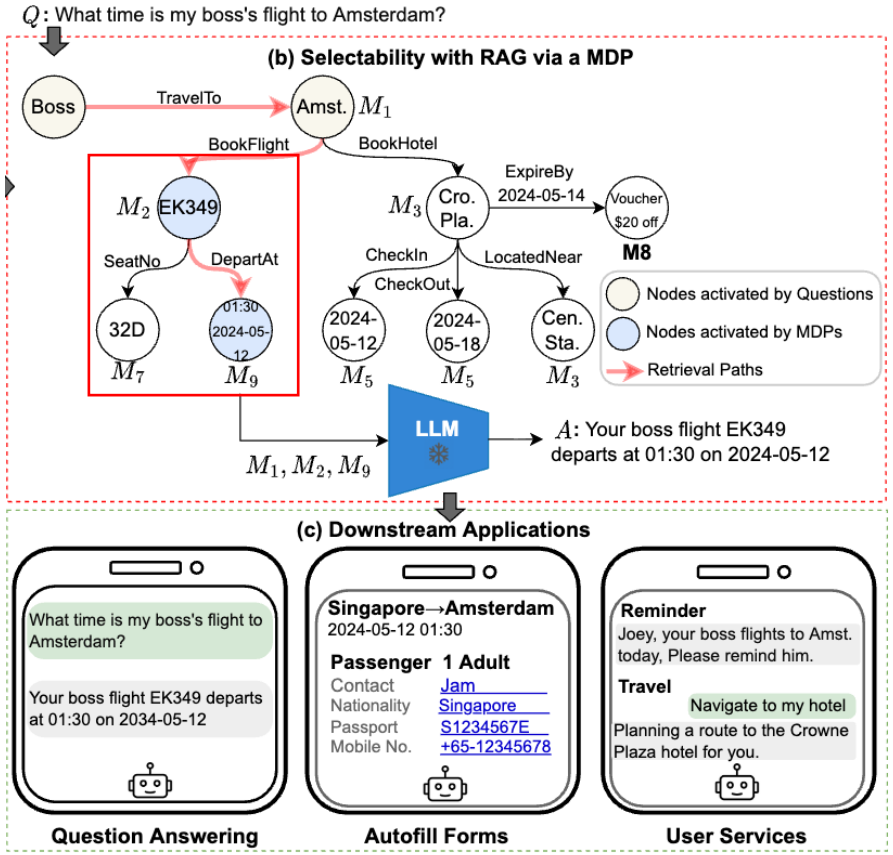
- Focusing on personalized selectability — adaptively retrieving the right combination of memories for each user's query.
- **EMG-RAG (EMNLP'24)**
  - **2. Retrieval Outline**
    - Activate Relevant Nodes



[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

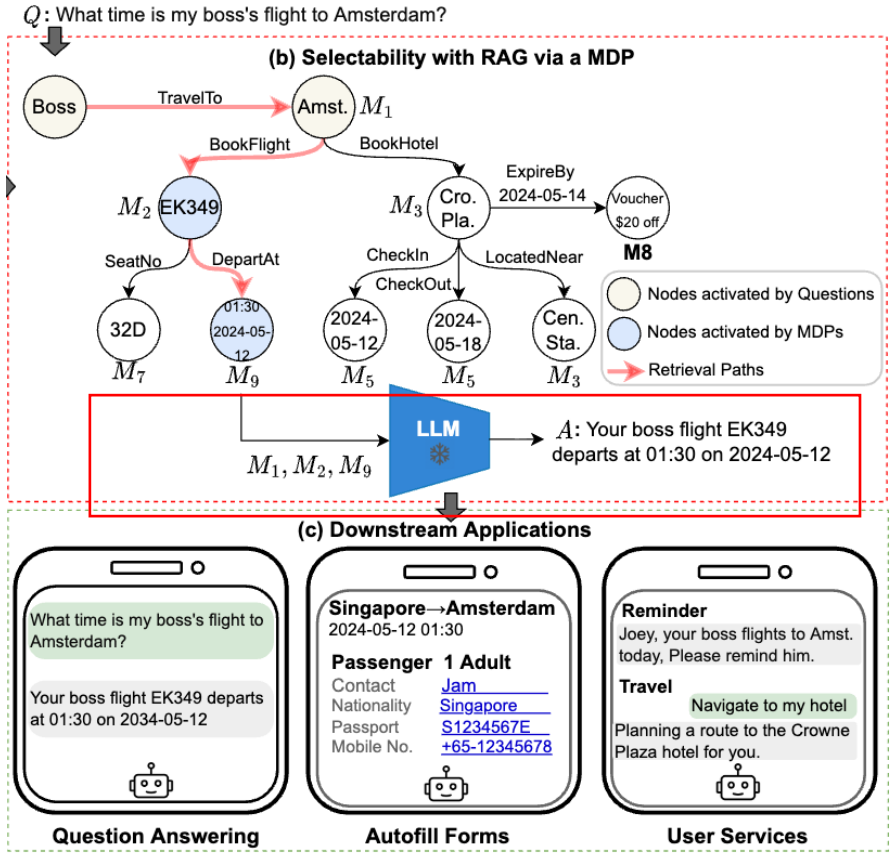
- Focusing on personalized selectability — adaptively retrieving the right combination of memories for each user's query.
- **EMG-RAG (EMNLP'24)**
  - **2. Retrieval Outline**
    - Activate Relevant Nodes
    - RL Agent Traverses the Graph



[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

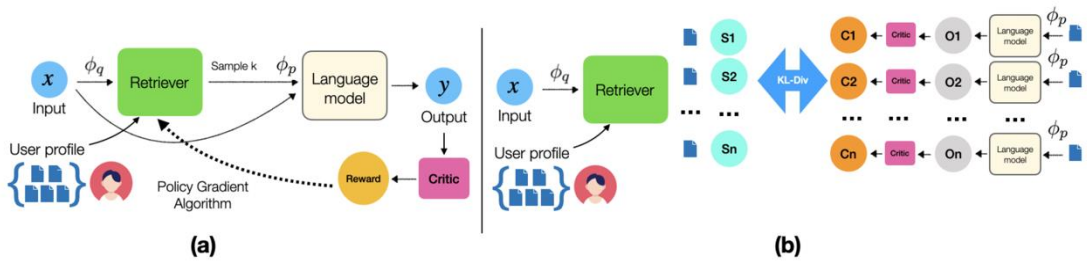
- Focusing on personalized selectability — adaptively retrieving the right combination of memories for each user's query.
- **EMG-RAG (EMNLP'24)**
  - **2. Retrieval Outline**
    - Activate Relevant Nodes
    - RL Agent Traverses the Graph
    - Reward Signal: change in LLM answer quality after including a memory
    - Training: Warm-Start (supervised fine-tuning) → Policy Gradient (reinforcement learning)



[1] Wang, Zheng, et al. "Crafting Personalized Agents through Retrieval-Augmented Generation on Editable Memory Graphs." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024.

Indexing

- ROPG enables personalized indexing by **optimizing the encoder** that builds the user profile index, replacing a general-purpose embedding model with one trained to prioritize personally useful documents.

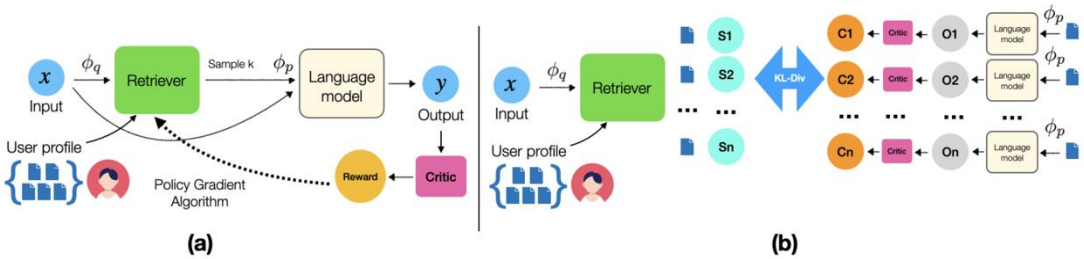


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- ROPG enables personalized indexing by **optimizing the encoder** that builds the user profile index, replacing a general-purpose embedding model with one trained to prioritize personally useful documents.

- **ROPG (SIGIR'24)**
  - **Problem:**
    - User profile documents are indexed using general-purpose encoders not trained for personalization
    - No document-level relevance labels exist

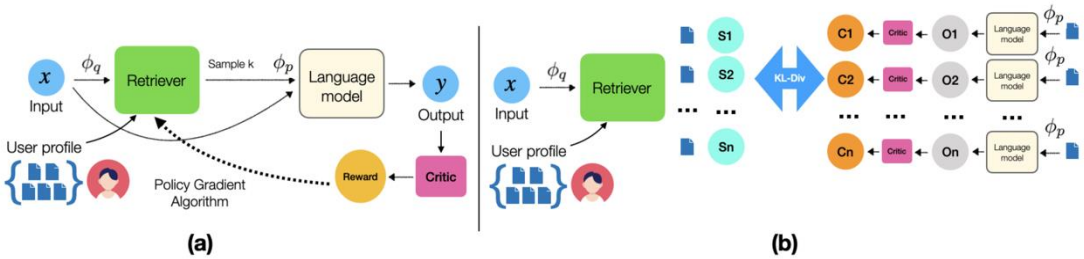


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- ROPG enables personalized indexing by **optimizing the encoder** that builds the user profile index, replacing a general-purpose embedding model with one trained to prioritize personally useful documents.

- **ROPG (SIGIR'24)**
  - **Problem:**
    - User profile documents are indexed using general-purpose encoders not trained for personalization
    - No document-level relevance labels exist
  - **Method:**
    - Uses policy gradient RL to optimize the retriever
    - Uses knowledge distillation from LLM
    - A retrieval selection model that adaptively selects the best-performing index encoder per input

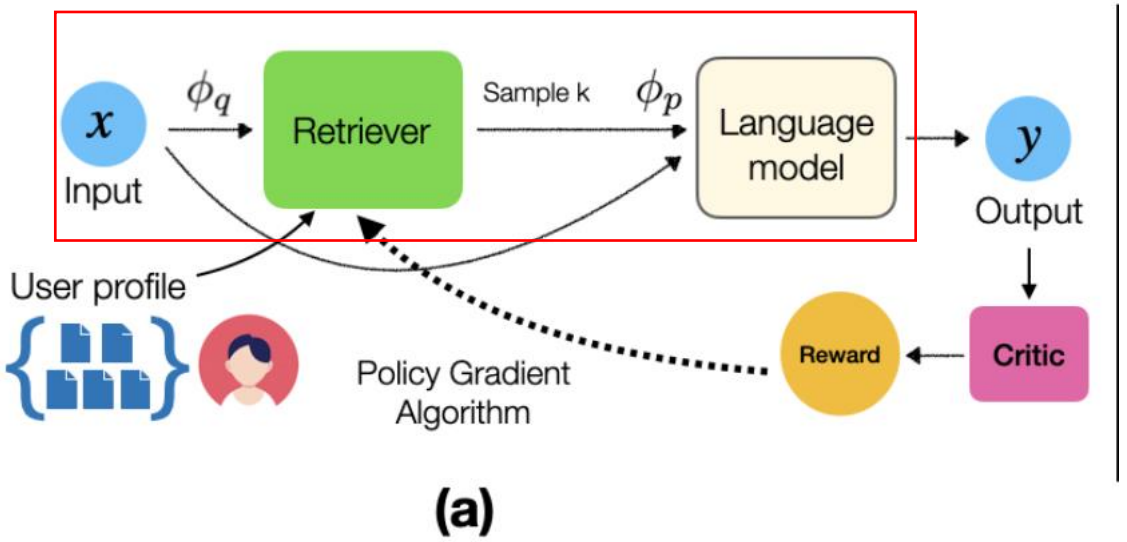


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

- **ROPG (SIGIR'24)**
  - **1. Retrieval Optimization via Reinforcement Learning**
    - The retrieval model is treated as a parameterized policy; selecting a document from the user profile is an action

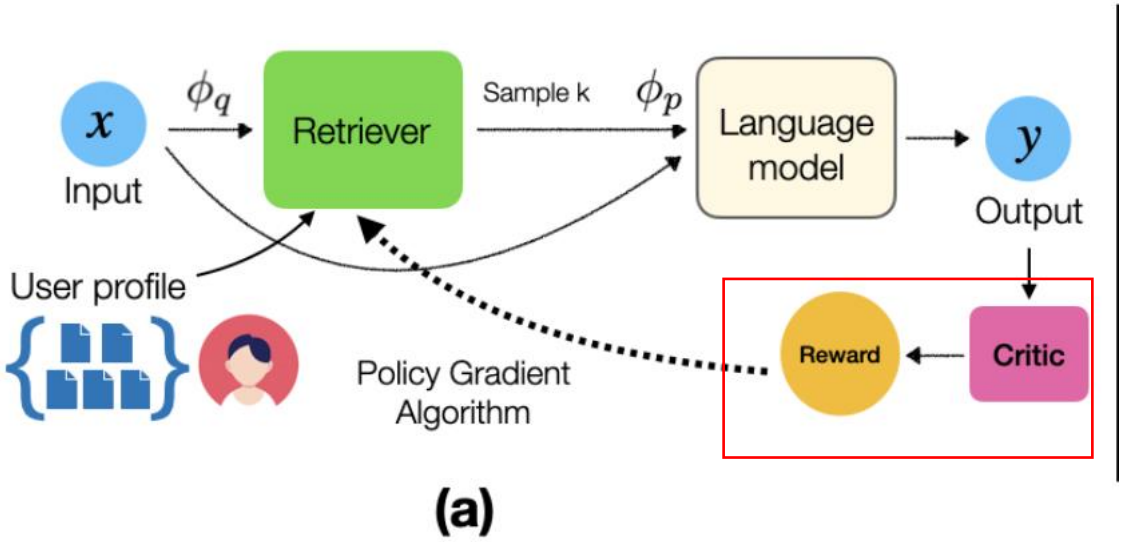


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

- **ROPG (SIGIR'24)**
  - **1. Retrieval Optimization via Reinforcement Learning**
    - The retrieval model is treated as a parameterized policy; selecting a document from the user profile is an action
    - **Reward** = improvement in LLM generation quality after retrieving a given document

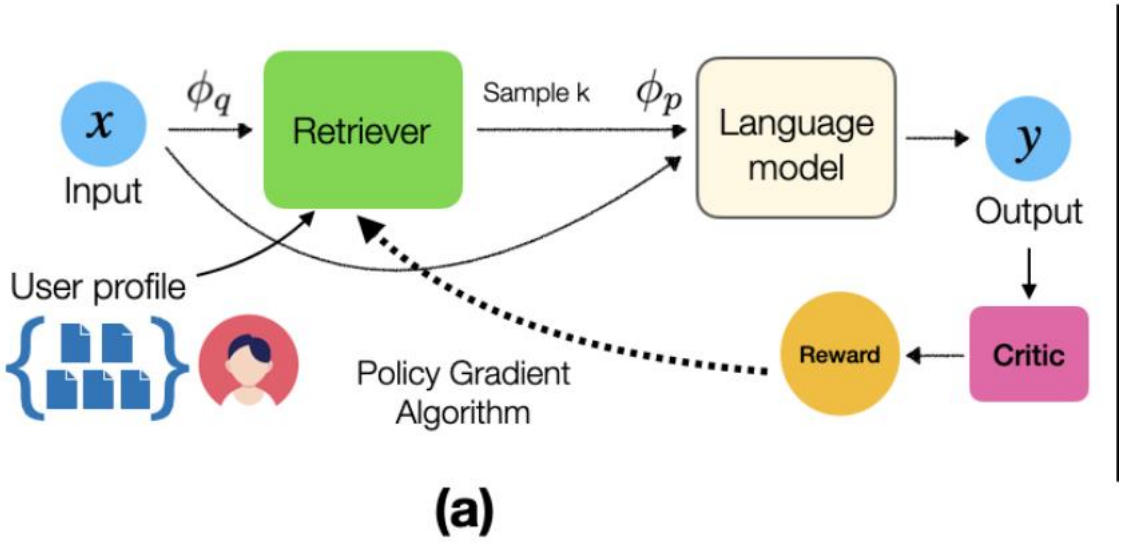


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

- **ROPG (SIGIR'24)**
  - **1. Retrieval Optimization via Reinforcement Learning**
    - The retrieval model is treated as a parameterized policy; selecting a document from the user profile is an action
    - **Reward** = improvement in LLM generation quality after retrieving a given document
    - Policy gradient updates push the retriever to prioritize documents that genuinely improve personalized generation



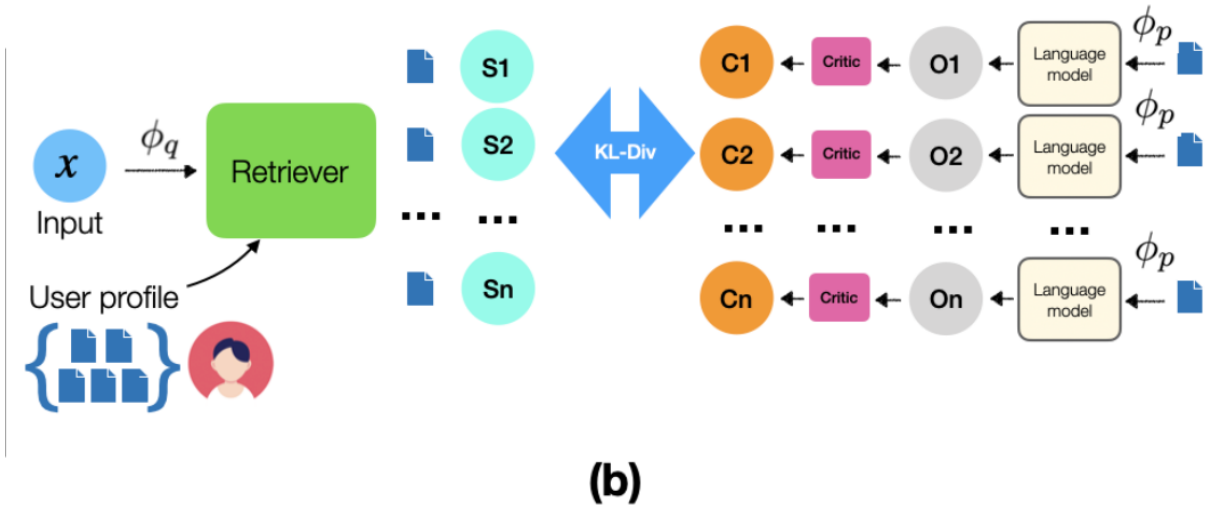
(a)

[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

- **ROPG (SIGIR'24)**
  - **2. Retrieval Optimization via Knowledge Distillation**

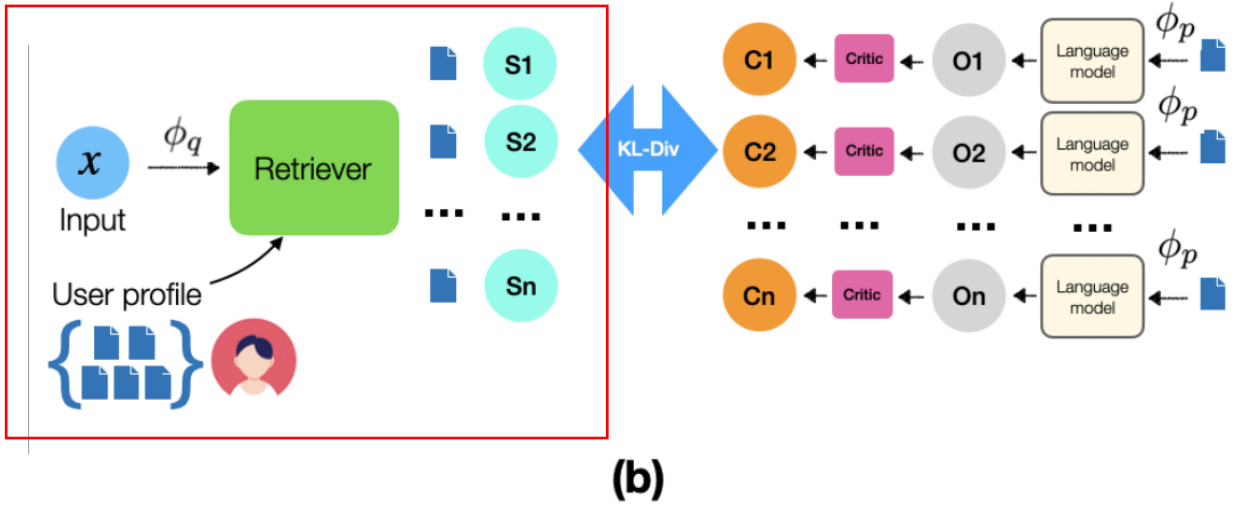


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

- **ROPG (SIGIR'24)**
  - **2. Retrieval Optimization via Knowledge Distillation**
    - LLM evaluates each profile document independently to produce a target score distribution



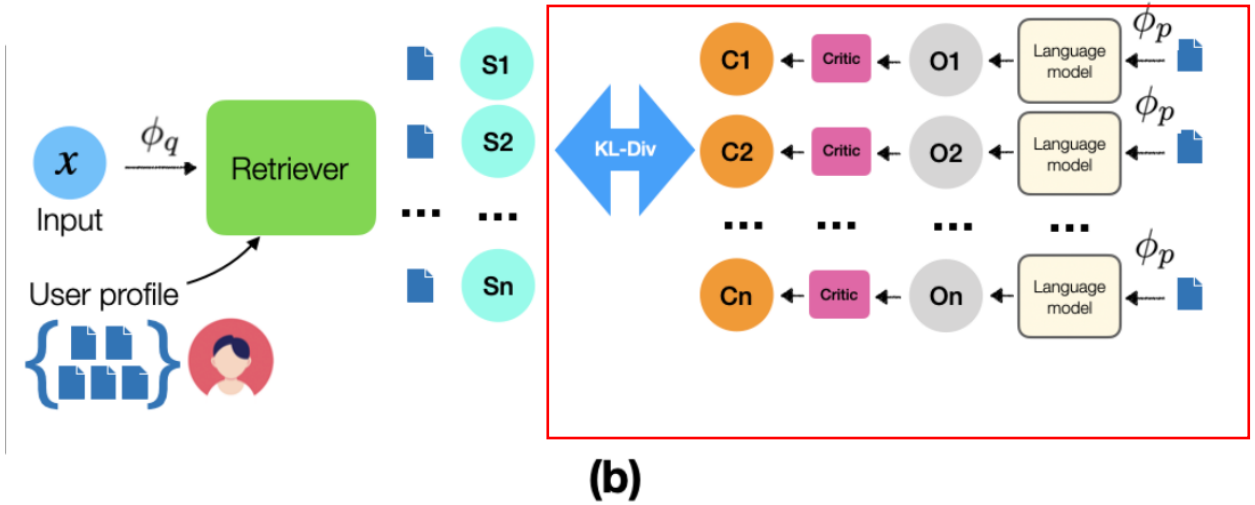
[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

➢Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

➢ ROPG (SIGIR'24)

- ➢ **2. Retrieval Optimization via Knowledge Distillation**
  - ➢ LLM evaluates each profile document independently to produce a target score distribution
  - ➢ Retriever is trained to align its scoring distribution with the LLM's document-level performance via KL-divergence loss

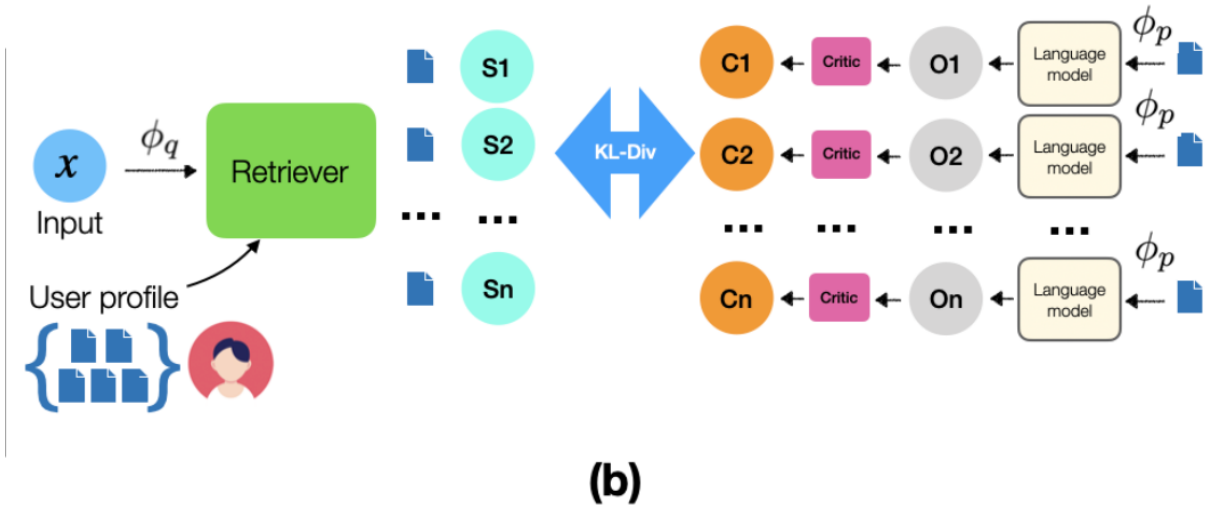


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

- **ROPG (SIGIR'24)**
  - **2. Retrieval Optimization via Knowledge Distillation**
    - LLM evaluates each profile document independently to produce a target score distribution
    - Retriever is trained to align its scoring distribution with the LLM's document-level performance via KL-divergence loss
    - More stable than RL; captures relative usefulness across all documents

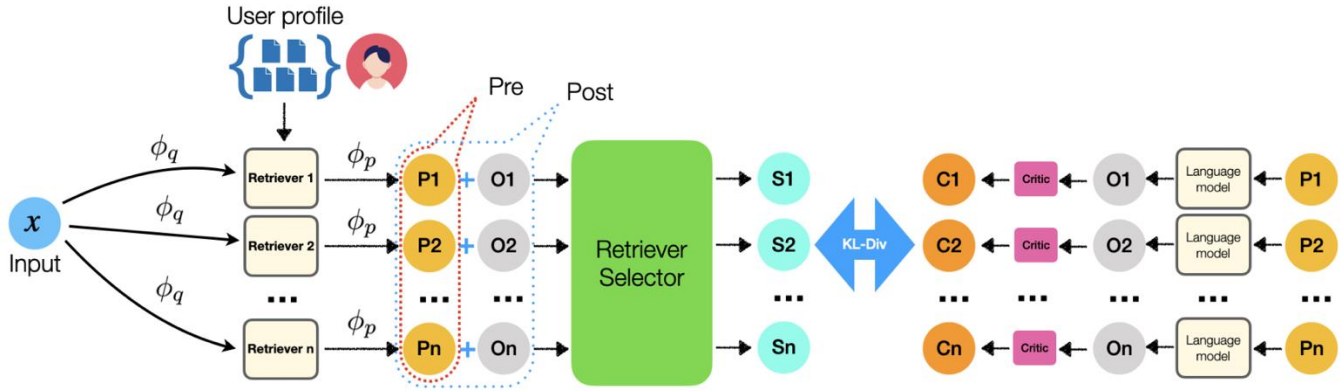


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Indexing

- Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

- **ROPG (SIGIR'24)**
  - **3. RSPG: Adaptive Retrieval Model Selection**
    - **Observation**: no single retriever wins across all personalization tasks

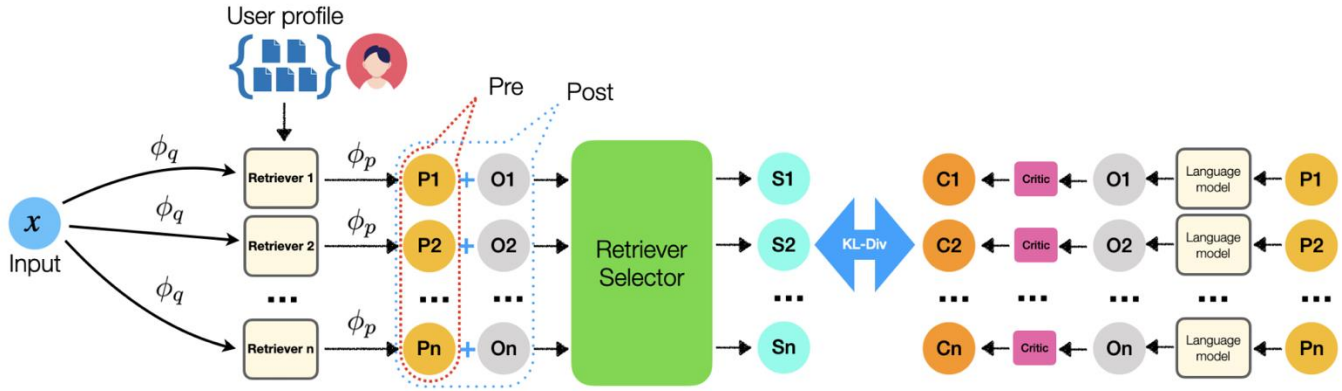


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024

Indexing

- Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

- **ROPG (SIGIR'24)**
  - **3. RSPG: Adaptive Retrieval Model Selection**
    - **Observation**: no single retriever wins across all personalization tasks
    - **Pre-generation**: selects the best retriever based on constructed prompts before LLM generation



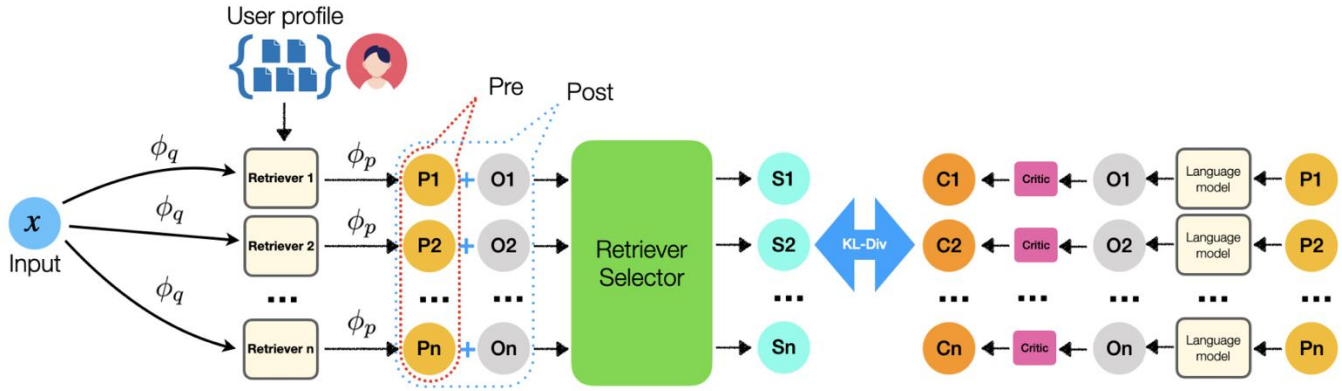
[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024

Indexing

➢ Focusing on personalized Retrieval: Optimizing the Retrieval Model with LLM Feedback

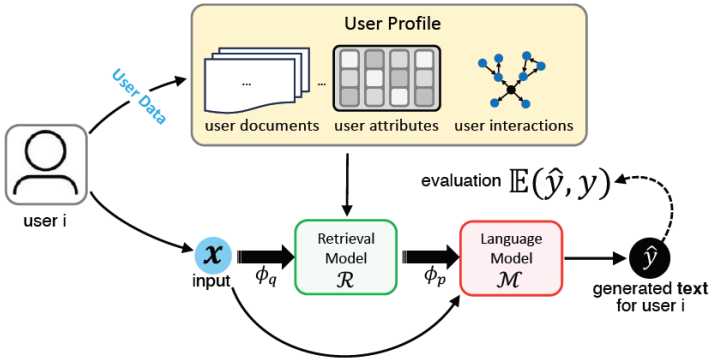
➢ **ROPG (SIGIR'24)**

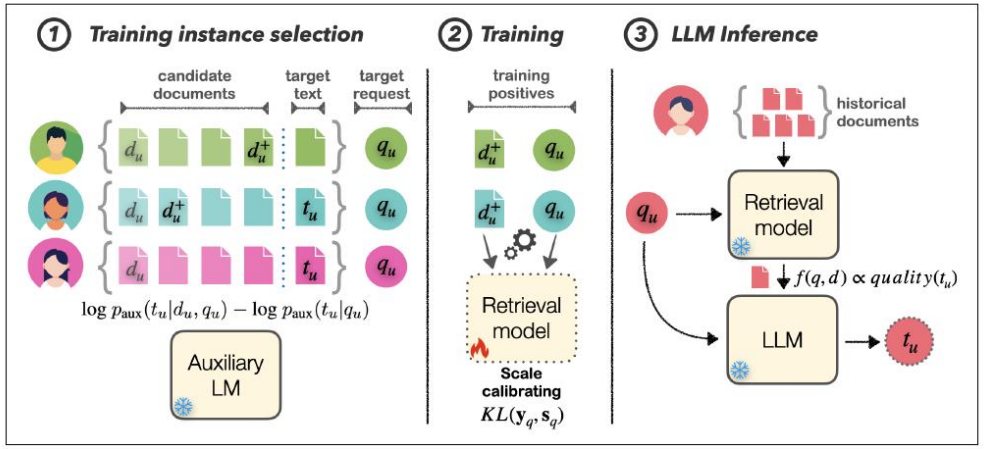
- ➢ **3. RSPG: Adaptive Retrieval Model Selection**
  - ➢ **Observation**: no single retriever wins across all personalization tasks
  - ➢ **Pre-generation**: selects the best retriever based on constructed prompts before LLM generation
  - ➢ **Post-generation**: makes selection after seeing LLM output, leveraging richer signal for better accuracy

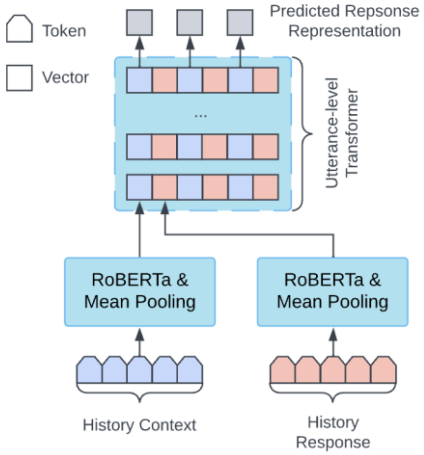


[1] Salemi, Alireza, Surya Kallumadi, and Hamed Zamani. "Optimization methods for personalizing large language models through retrieval augmentation." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024

Other Indexing Works







- Overview of PGraphRAG [1]
- Overview of PEARL[2]
- Overview of RECAP [3]

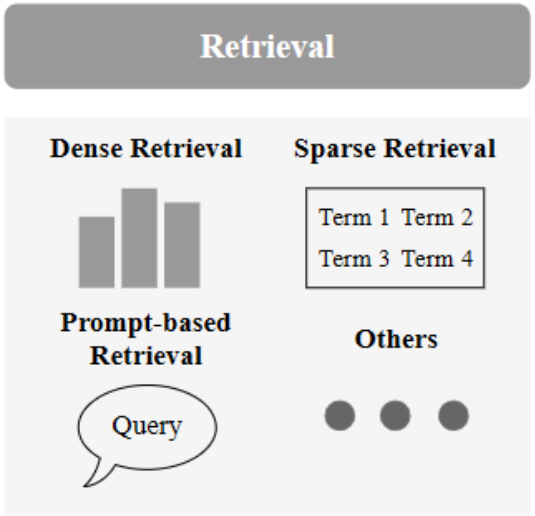
[1] Au, Steven, et al. "Personalized Graph-Based Retrieval for Large Language Models." *arXiv preprint arXiv:2501.02157* (2025).
[2] Mysore, Sheshera, et al. "Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers." arXiv preprint arXiv:2311.09180 (2023).
[3]] Liu, Shuai, et al. "RECAP: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.

Retrieval

- Motivation
  - ***Existing retrieval systems lack the ability to adapt to user-specific preferences and contexts, often leading to generalized and less relevant results.***

- Personalized Retrieval Definition
  - Personalized retrieval integrates user-specific information such as preferences, historical interactions, and contextual data to deliver more relevant and tailored search results.

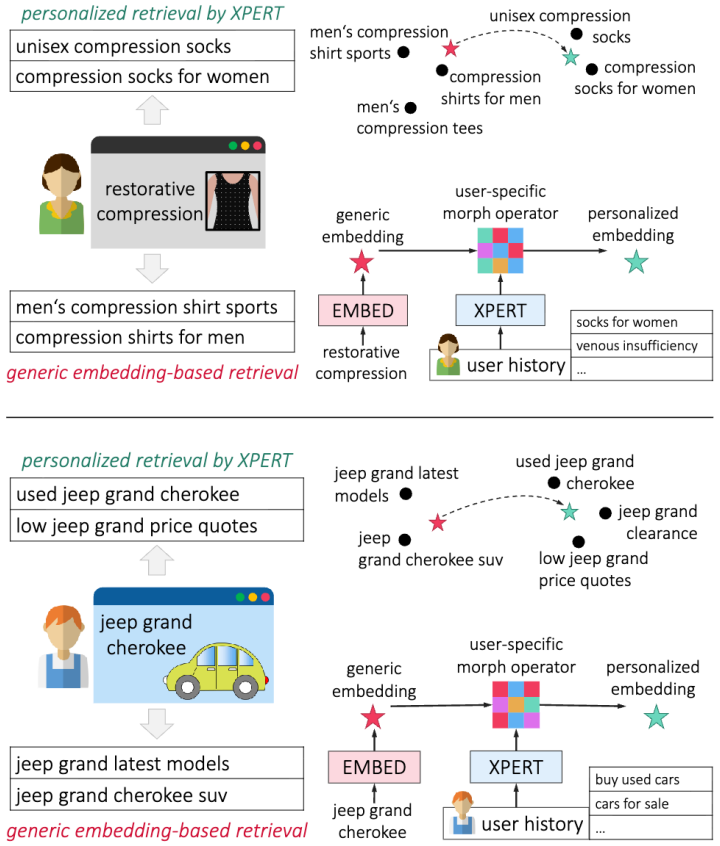
- Methods can be divided to:
  - Dense Retrieval
  - Sparse Retrieval
  - Prompt-Based Retrieval



Retrieval - Dense Retrieval

- **Dense retrieval** methods often achieve personalization by **encoding user preferences, context, or interactions into query or document embeddings**.

- **XPERT (SIGIR'23)**
  - **Problem:**
    - Two-sided personalization (event/item) requires a separate MIPS index per user — infeasible at million-item scale
  - **Method:**
    - Morph Operators
    - 3-Segment Pipeline

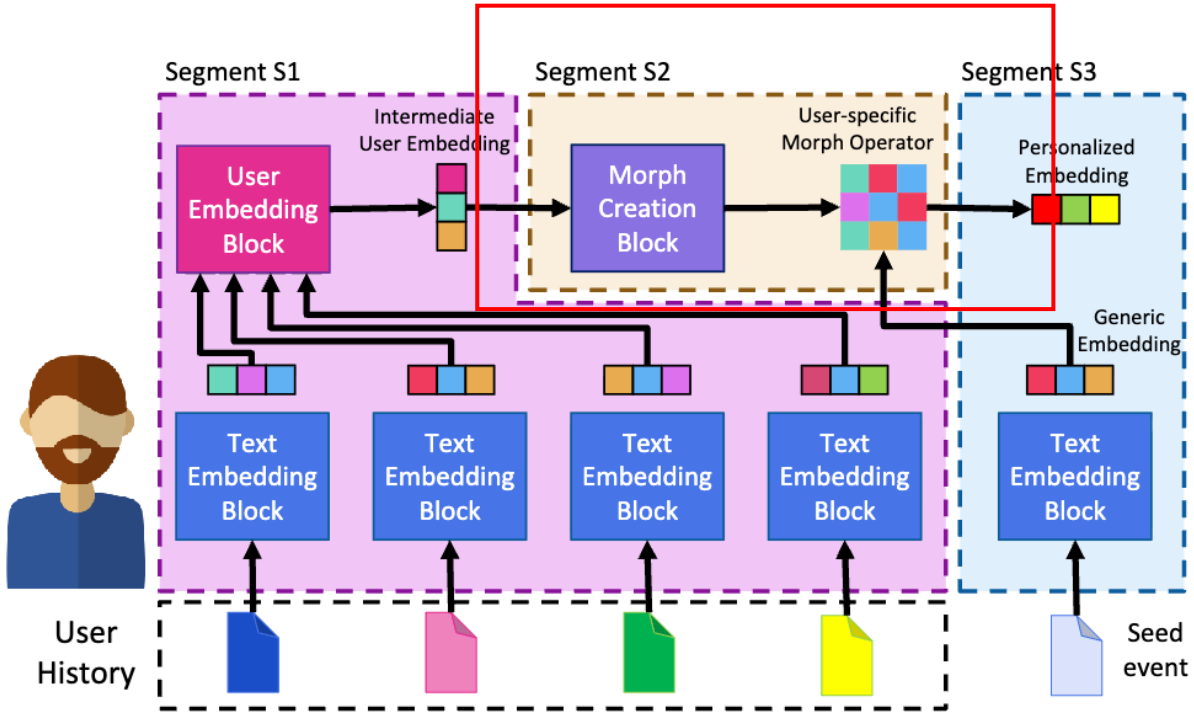


[1] Vemuri H, Agrawal S, Mittal S, et al. Personalized retrieval over millions of items[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 1014-1022.

Retrieval - Dense Retrieval

- XPERT achieves personalized retrieval by learning a **user-specific morph operator** from browsing history to transform generic event embeddings, enabling scalable two-sided personalization over a **single shared MIPS index**.

- **XPERT (SIGIR'23)**
  - **1. Morph Operator**
    - Acts as a transformation matrix
    - Rotates the generic event embedding toward the user's preference space
    - Item embeddings remain generic → single shared MIPS index is preserved

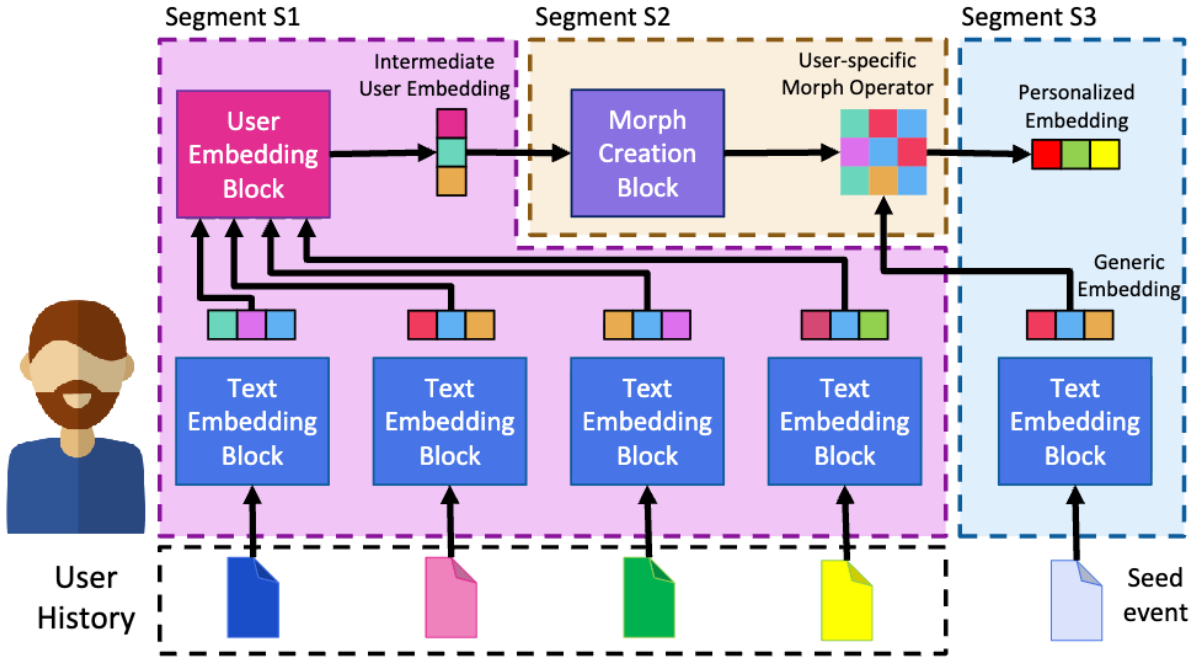


[1] Vemuri H, Agrawal S, Mittal S, et al. Personalized retrieval over millions of items[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 1014-1022.

Retrieval - Dense Retrieval

- XPERT achieves personalized retrieval by learning a **user-specific morph operator** from browsing history to transform generic event embeddings, enabling scalable two-sided personalization over a **single shared MIPS index**.

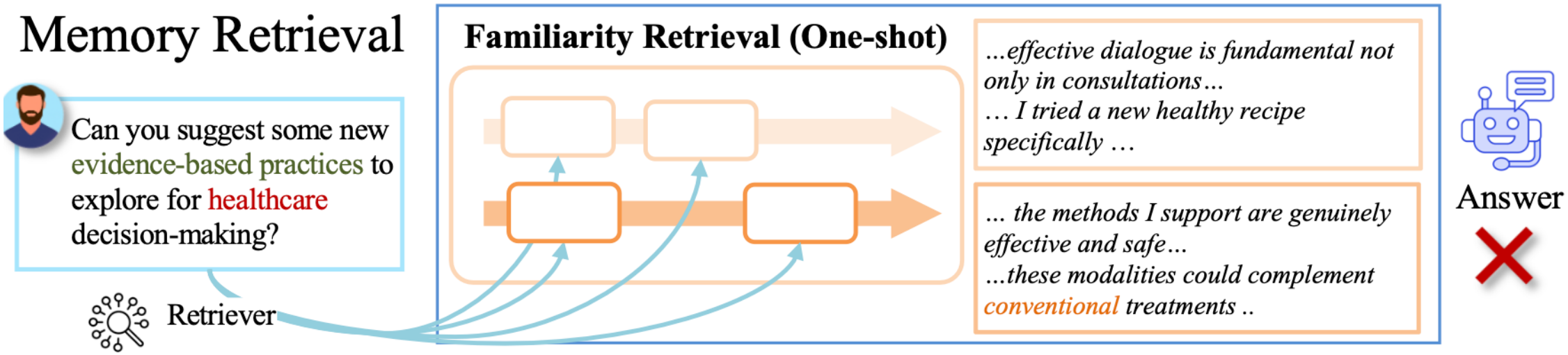
- **XPERT (SIGIR'23)**
  - **2. Segment Prediction Pipeline**
    - S1 — History Aggregation: Encodes user history events via a Transformer → user embedding
    - S2 — Morph Generation: User embedding fed into a feed-forward layer → morph operator matrix
    - S3 — Retrieval: Morph operator applied to seed event embeddings → MIPS query against item index



[1] Vemuri H, Agrawal S, Mittal S, et al. Personalized retrieval over millions of items[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 1014-1022.

Retrieval - Dense Retrieval

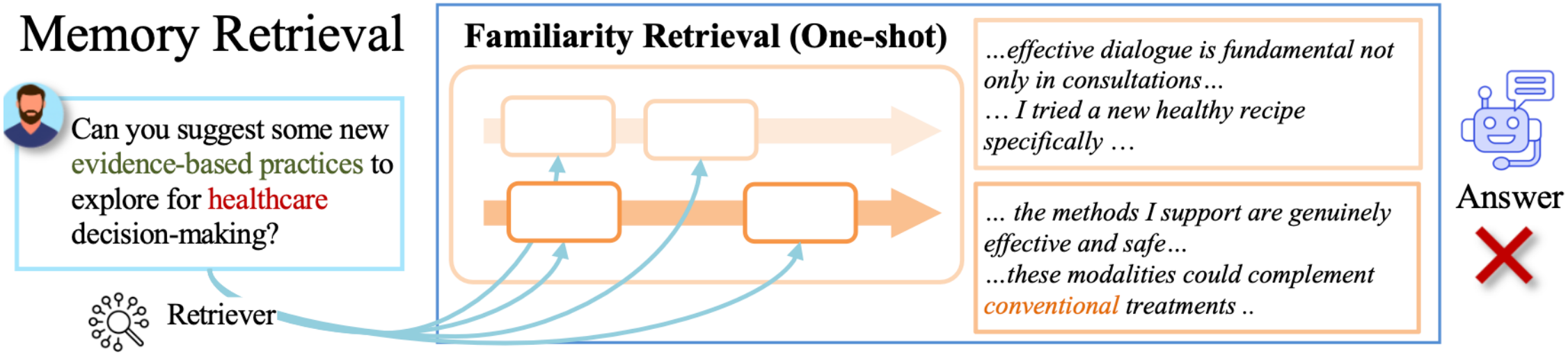
- **Problem:**
  - Existing retrieval methods mainly rely on one-shot similarity search, which captures only surface-level matches.
  - Complex personalized queries often require recovering contextual evidence scattered across user memories
- **RF-Mem (Ours ICLR26)**



[1] Zhang, Yingyi, et al. "Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval." The Fourteenth International Conference on Learning Representations.

Retrieval - Dense Retrieval

- Cognitive science suggests that human memory operates through two complementary processes:
  - Familiarity — fast recognition
  - Recollection — deliberate contextual reconstruction

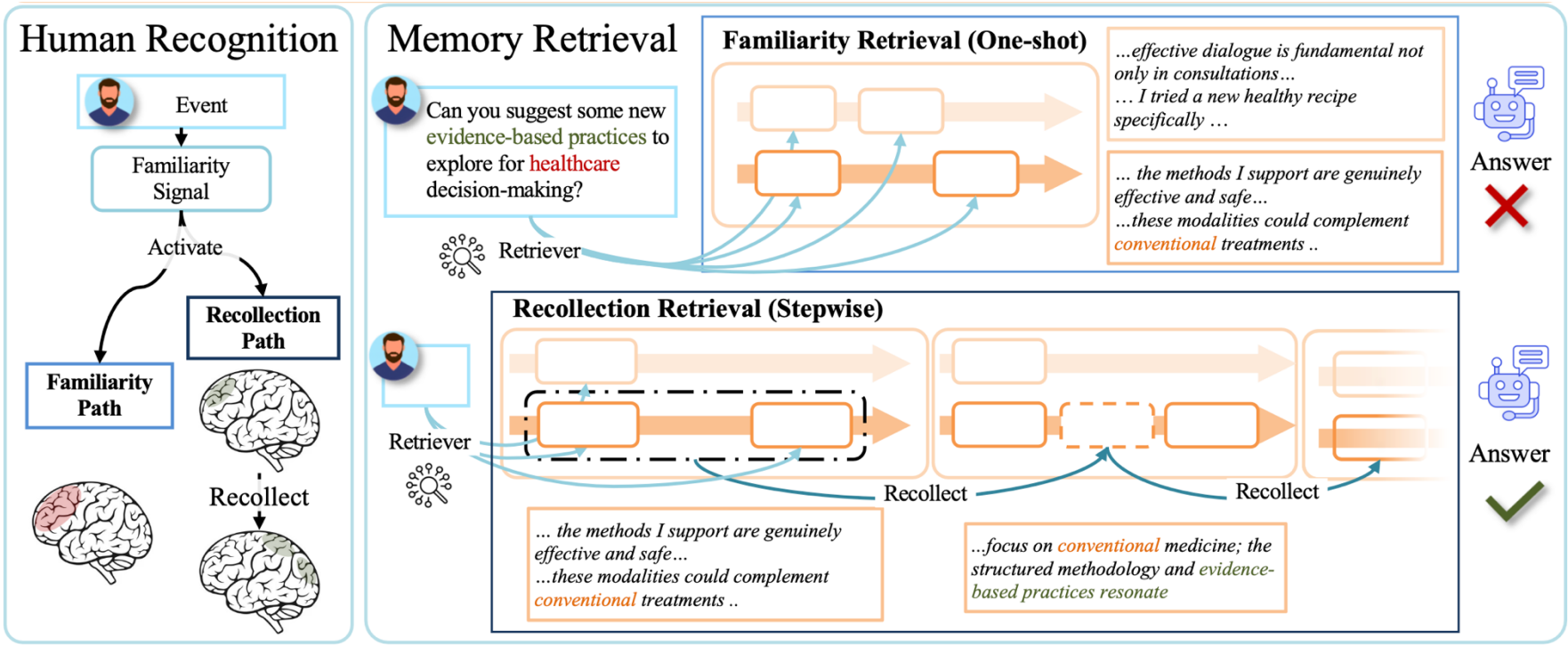


- Current systems lack both the ability to perform recollection retrieval and mechanisms to adaptively switch between the dual retrieval paths, leading to either insufficient recall or the inclusion of noise.

[1] Zhang, Yingyi, et al. "Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval." The Fourteenth International Conference on Learning Representations.

Retrieval - Dense Retrieval

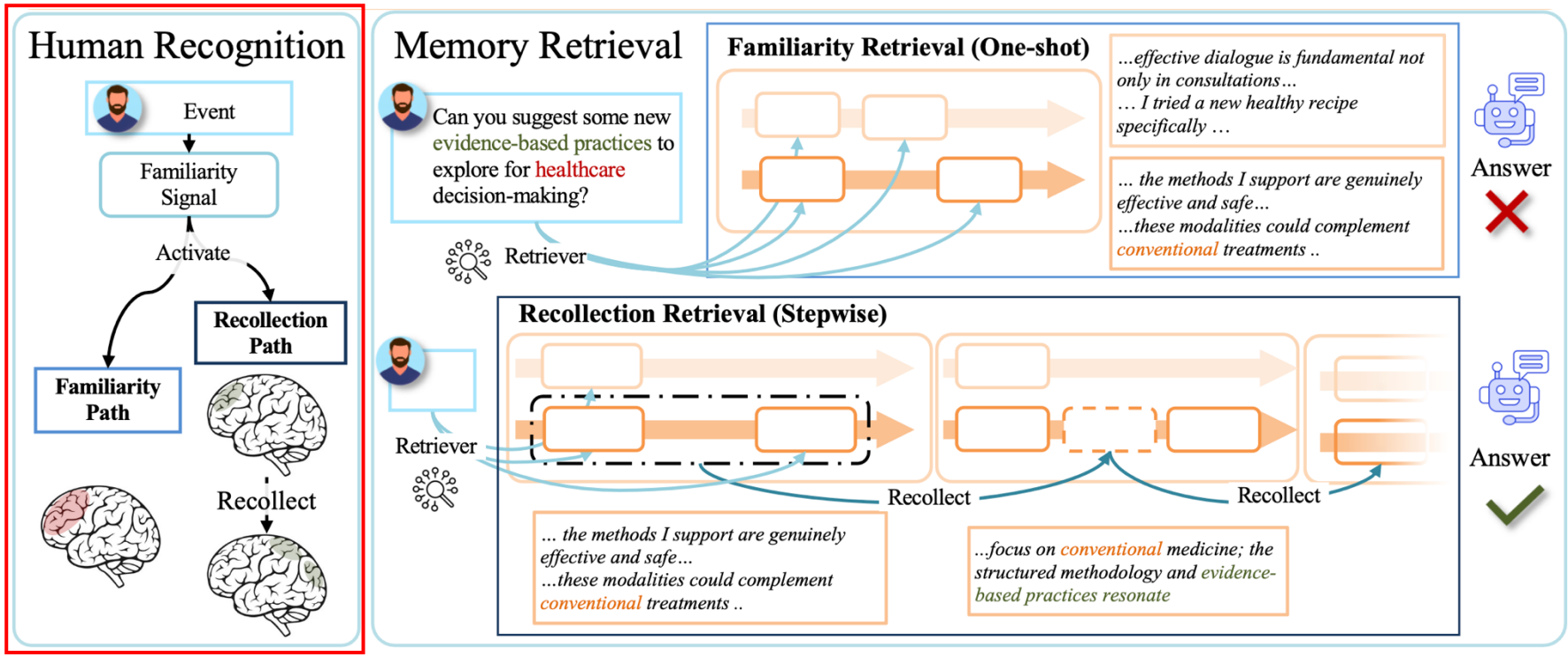
- Core Idea:
  - Propose RF-Mem (Recollection–Familiarity Memory Retrieval), a familiarity uncertainty-guided dual-path memory retriever.



[1] Zhang, Yingyi, et al. "Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval." The Fourteenth International Conference on Learning Representations.

Retrieval - Dense Retrieval

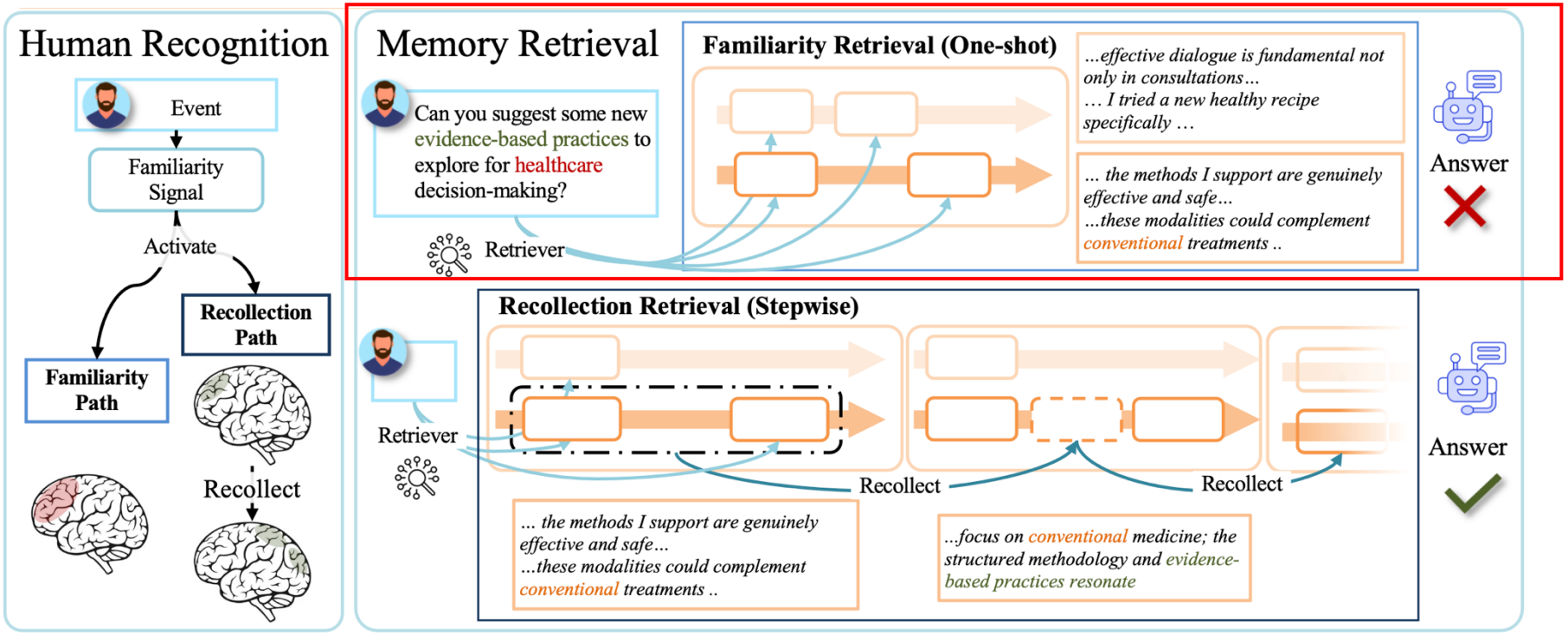
- Core Idea:
  - Propose RF-Mem (Recollection–Familiarity Memory Retrieval), a familiarity uncertainty-guided dual-path memory retriever.



[1] Zhang, Yingyi, et al. "Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval." The Fourteenth International Conference on Learning Representations.

Retrieval - Dense Retrieval

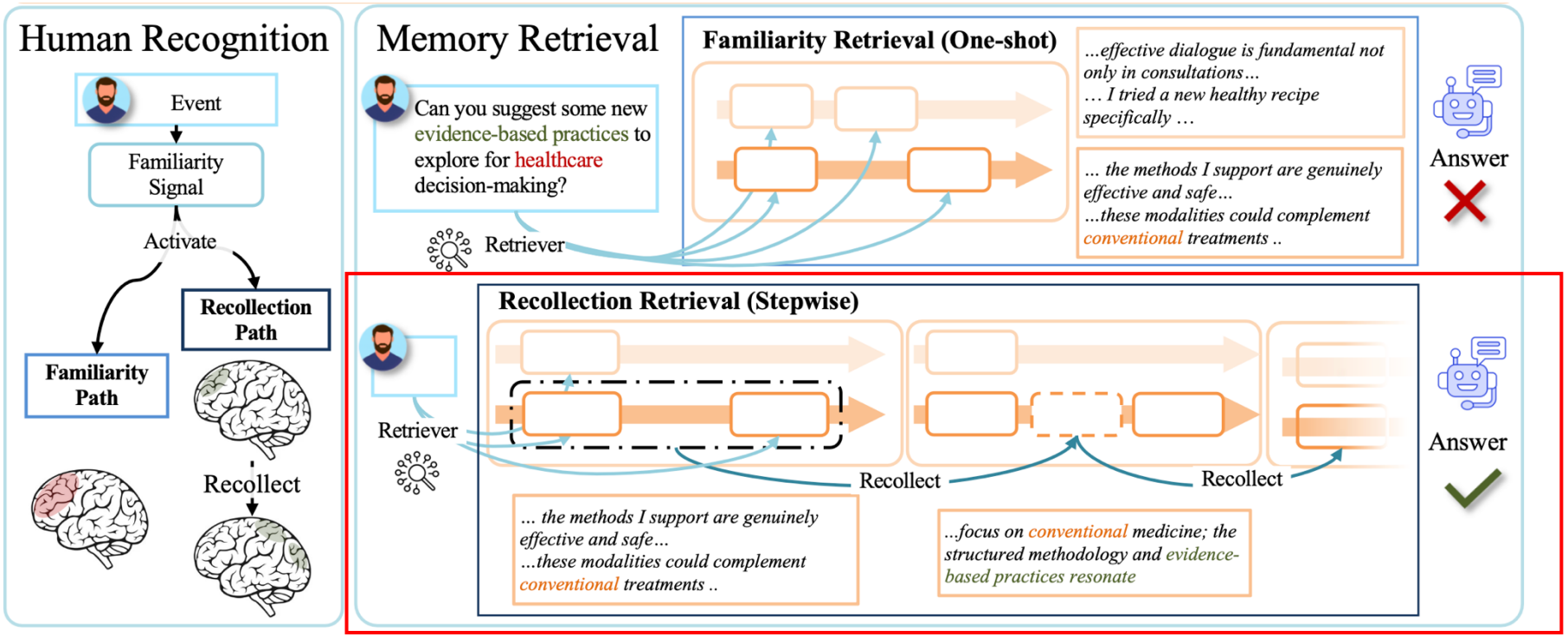
- Core Idea:
  - Propose RF-Mem (Recollection–Familiarity Memory Retrieval), a familiarity uncertainty-guided dual-path memory retriever.



[1] Zhang, Yingyi, et al. "Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval." The Fourteenth International Conference on Learning Representations.

Retrieval - Dense Retrieval

- Core Idea:
  - Propose RF-Mem (Recollection–Familiarity Memory Retrieval), a familiarity uncertainty-guided dual-path memory retriever.

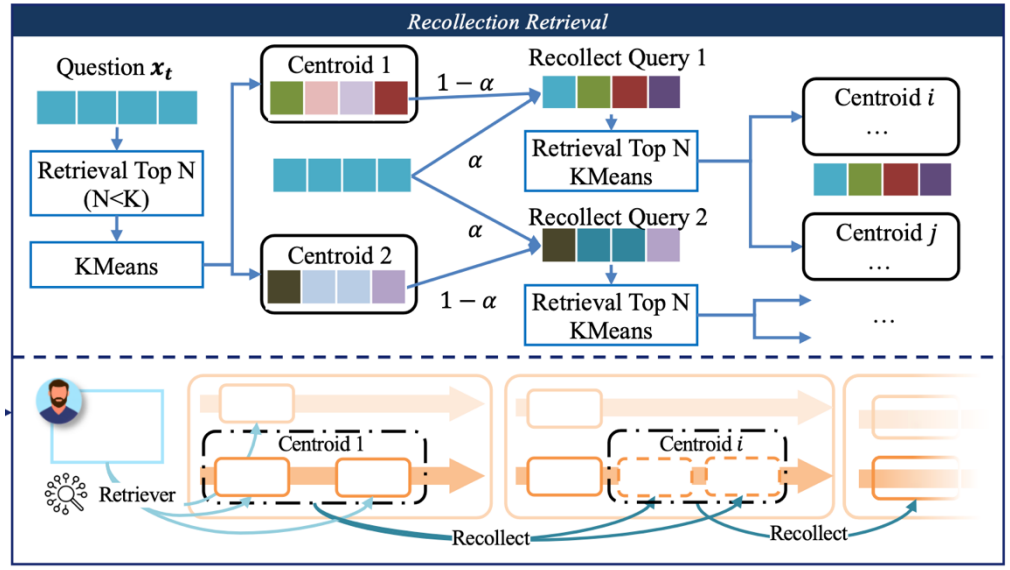


[1] Zhang, Yingyi, et al. "Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval." The Fourteenth International Conference on Learning Representations.

Retrieval - Dense Retrieval

RF-Mem Framework RF-Mem first performs a probe retrieval to estimate a familiarity signal using the mean similarity score and entropy of retrieved candidates.

- If the familiarity signal is strong, the system directly performs Familiarity retrieval, returning the top-K memories via similarity search.
- Otherwise, RF-Mem activates the Recollection path, which iteratively expands evidence by clustering retrieved memories and generating new recollect queries through query–centroid mixing.
  - 1. Candidate Memory Retrieval.
  - 2. Relevant Memory Clustering.
  - 3. Recollect Queries Generation via α-mix.
  - 4. Retrieve-Cluster-Mix Loop.
  - 5. Stop



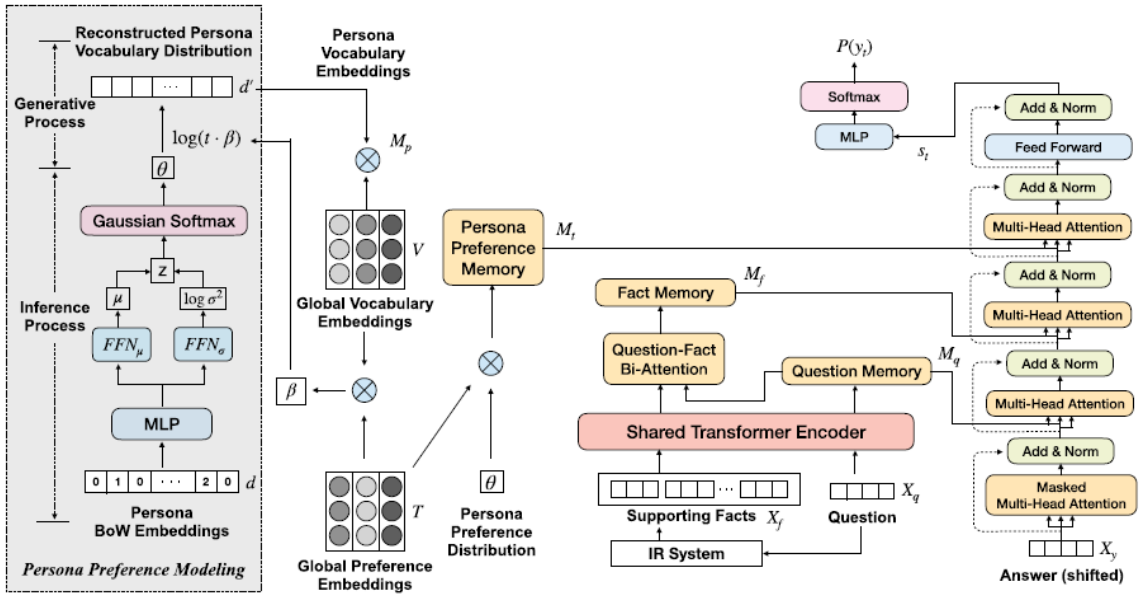
[1] Zhang, Yingyi, et al. "Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval." The Fourteenth International Conference on Learning Representations.

Retrieval - Sparse Retrieval

➢ **Sparse retrieval** methods apply personalization by **assigning higher weights to terms that reflect user preferences**.

➢ **PAGE (TOIS'22):**

- ➢ **Problem:**
  - ➢ Existing PQA methods generate a single unified answer for all users, ignoring individual preferences
- ➢ **Method:**
  - ➢ Persona History Incorporation (PHI)
  - ➢ Persona Preference Modeling (PPM)
  - ➢ Persona Information Summarizer (PIS)

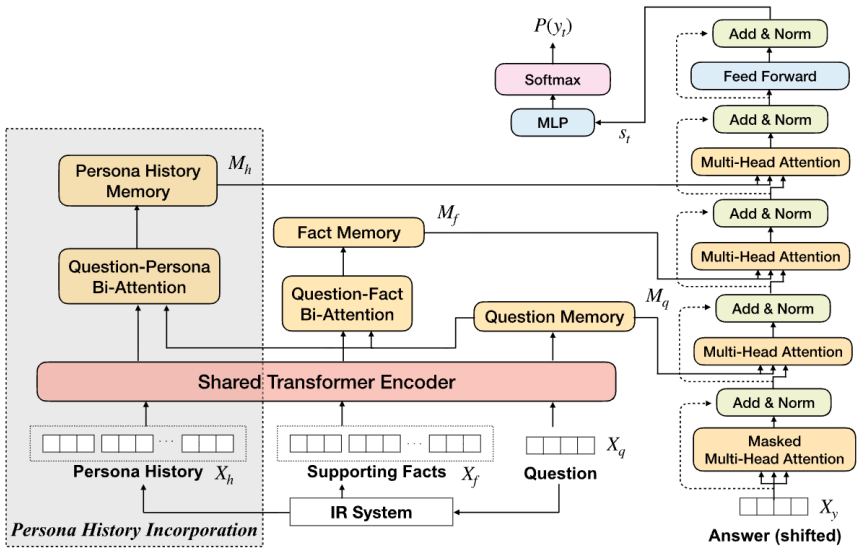


[1] Deng Y, Li Y, Zhang W, et al. Toward personalized answer generation in e-commerce via multi-perspective preference modeling[J]. ACM Transactions on Information Systems (TOIS), 2022, 40(4): 1-28.

Retrieval - Sparse Retrieval

- Personalization is achieved by retrieving and aggregating user-specific signals at three levels: knowledge, aspect, and vocabulary.

- **PAGE (TOIS'22):**
  - **1. Persona History Incorporation (PHI)**
    - Retrieves question-relevant snippets from user historical UGC (reviews, answers, questions) via BM25
    - Encodes retrieved snippets using a shared Self-Attentive Encoder and Bi-Attention mechanism
    - Produces a Persona History Memory Vector, injected into the decoder as an additional attention layer

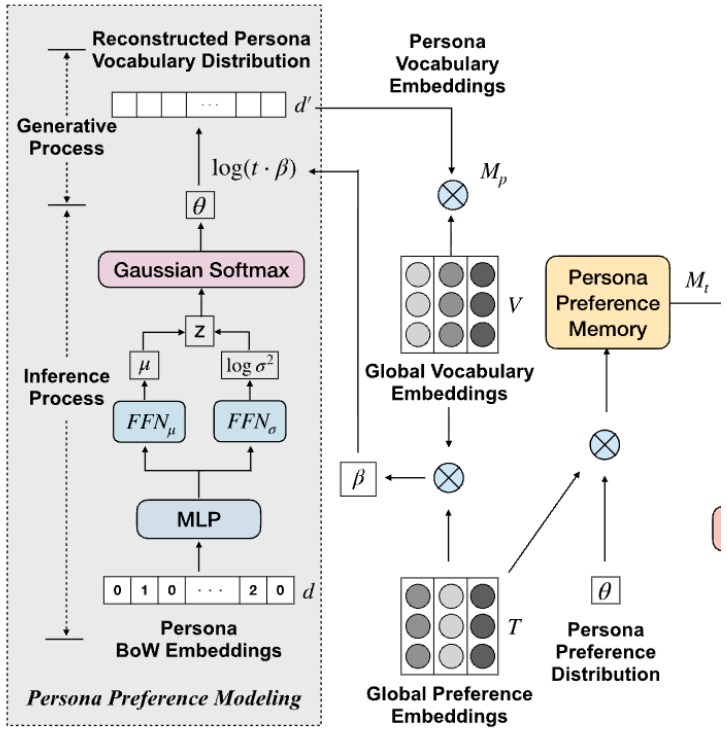


[1] Deng Y, Li Y, Zhang W, et al. Toward personalized answer generation in e-commerce via multi-perspective preference modeling[J]. ACM Transactions on Information Systems (TOIS), 2022, 40(4): 1-28.

Retrieval - Sparse Retrieval

- Personalization is achieved by retrieving and aggregating user-specific signals at three levels: knowledge, aspect, and vocabulary.

- **PAGE (TOIS'22):**
  - **2. Persona Preference Modeling (PPM)**
    - Takes user historical UGC converted into Bag-of-Words representation as input
    - Applies Gaussian Softmax variational inference to learn a latent topic distribution over user preferences
    - Outputs two components: a Persona Preference Memory (aspect-level) and a Persona Vocabulary Vector (vocabulary-level)
    - Both are incorporated into the decoder as additional attention layers

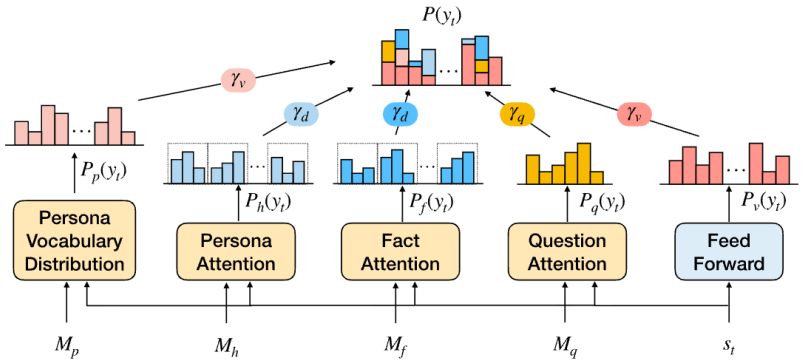


[1] Deng Y, Li Y, Zhang W, et al. Toward personalized answer generation in e-commerce via multi-perspective preference modeling[J]. ACM Transactions on Information Systems (TOIS), 2022, 40(4): 1-28.

Retrieval - Sparse Retrieval

- Personalization is achieved by retrieving and aggregating user-specific signals at three levels: knowledge, aspect, and vocabulary.

- **PAGE (TOIS'22):**
  - **3. Persona Information Summarizer (PIS)**
    - Computes attention weights over five word sources: question, supporting facts, user history, preference vocabulary, and base vocabulary
    - Generates a context vector for each source independently
    - Combines all five probability distributions via a learned multi-view scalar to produce the final word distribution

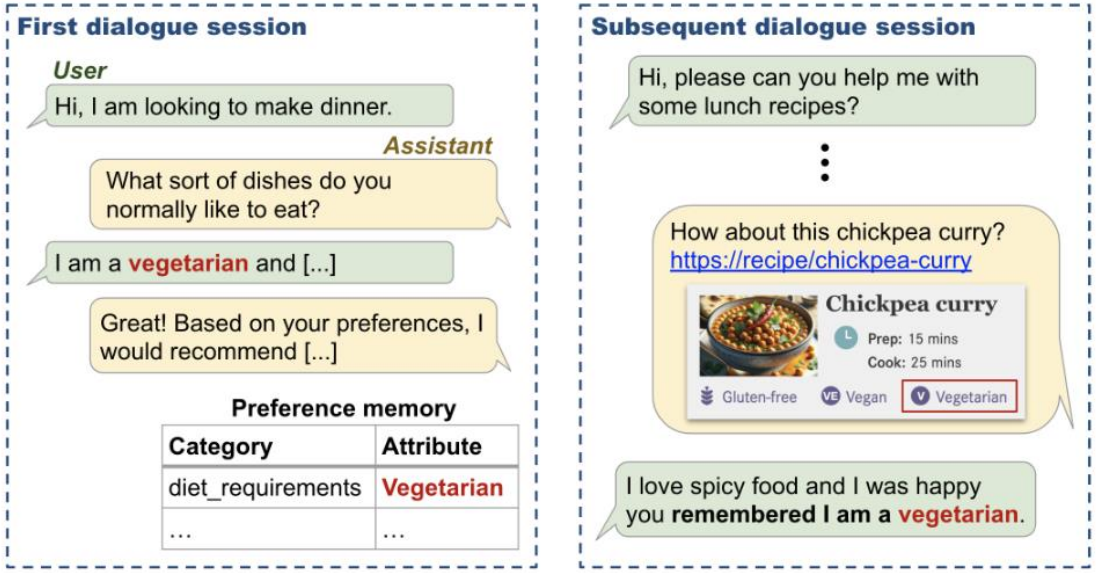


[1] Deng Y, Li Y, Zhang W, et al. Toward personalized answer generation in e-commerce via multi-perspective preference modeling[J]. ACM Transactions on Information Systems (TOIS), 2022, 40(4): 1-28.

Retrieval - Prompt-based Retrieval

- Personalization in **prompt-based retrieval** is achieved by **crafting user-specific prompts that embed historical interactions, preferences, or instructions**.

- **LAPS (SIGIR '24)**
  - **Problem:**
    - Large-scale human-written personalized dialogue datasets are unavailable, blocking development of personal conversational search systems
    - Fully synthetic LLM-generated dialogues lack diversity and do not reflect real user preferences
  - **Method:**
    - LLM-Augmented Self-Dialogue
    - Dialogue Act Classification
    - Preference Extraction & Memory

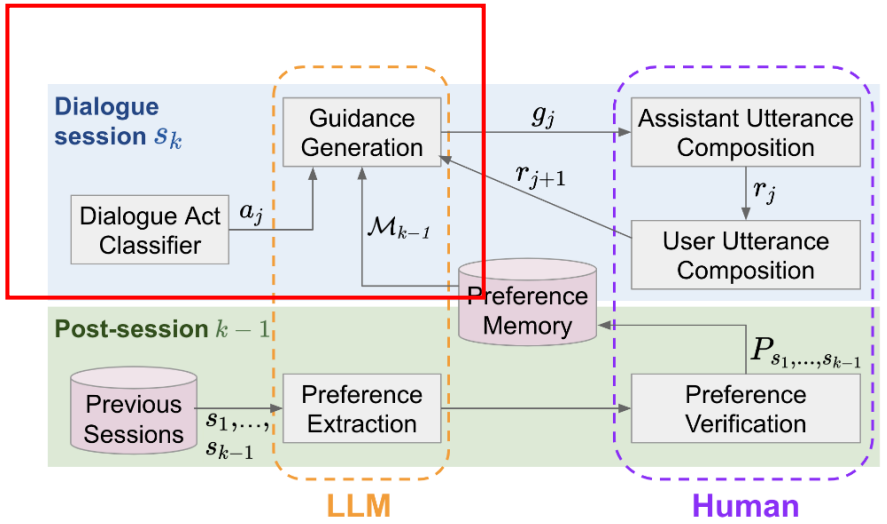


[1] Joko, Hideaki, et al. "Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Retrieval - Prompt-based Retrieval

- LAPS personalizes retrieval by extracting user preferences from multi-session dialogues and storing them in a structured preference memory

- **LAPS (SIGIR '24)**
  - **1. Dialogue Act Classification**
    - A dialogue act classifier prompts GPT-3.5 to determine the next action based on dialogue history and the previous act
    - The determined dialogue act is passed to a guidance generation function
    - GPT-3.5 generates step-by-step personalized guidance for the human worker, taking dialogue history, current dialogue act, and preference memory as input

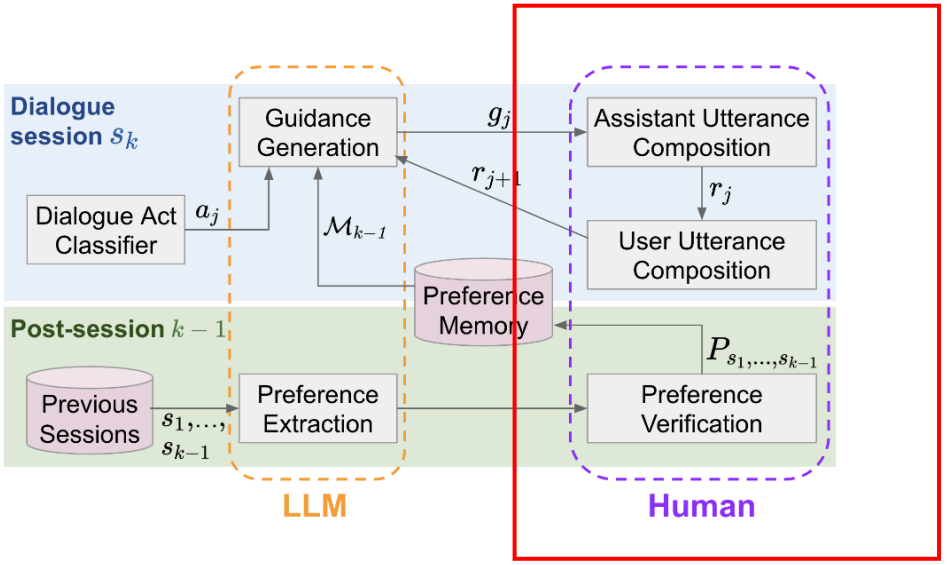


[1] Joko, Hideaki, et al. "Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Retrieval - Prompt-based Retrieval

- LAPS personalizes retrieval by extracting user preferences from multi-session dialogues and storing them in a structured preference memory

- **LAPS (SIGIR '24)**
  - **2. Utterance Composition**
    - A single crowd worker plays both user and assistant roles via dialogue self-play
    - For the assistant role, the worker composes responses based on the LLM-generated guidance
    - For the user role, the worker freely expresses personal preferences and reacts to recommendations

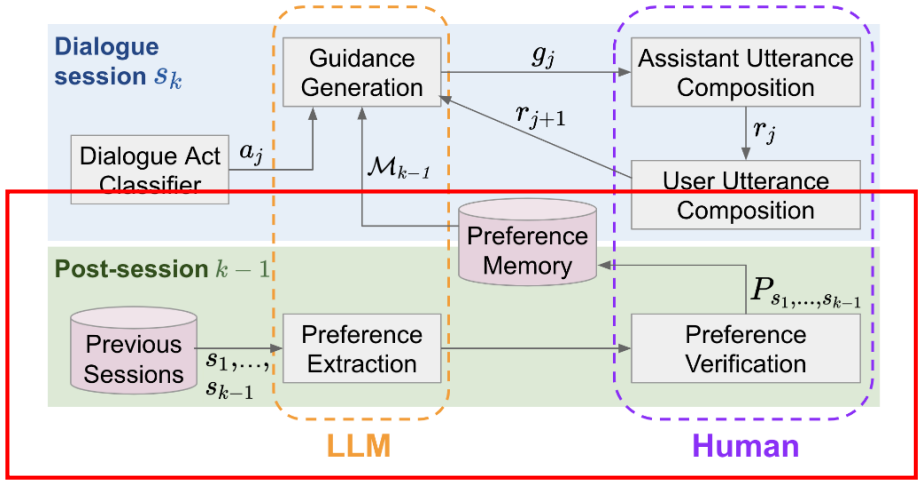


[1] Joko, Hideaki, et al. "Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Retrieval - Prompt-based Retrieval

- LAPS personalizes retrieval by extracting user preferences from multi-session dialogues and storing them in a structured preference memory

- **LAPS (SIGIR '24)**
  - **3. Preference Extraction**
    - After each session, GPT-4 extracts user preferences from the dialogue as category-preference pairs (e.g., diet_requirements: Vegetarian)
    - The human worker verifies and corrects the extracted preferences before they are stored
    - Extracted preferences are accumulated across sessions into a structured preference memory



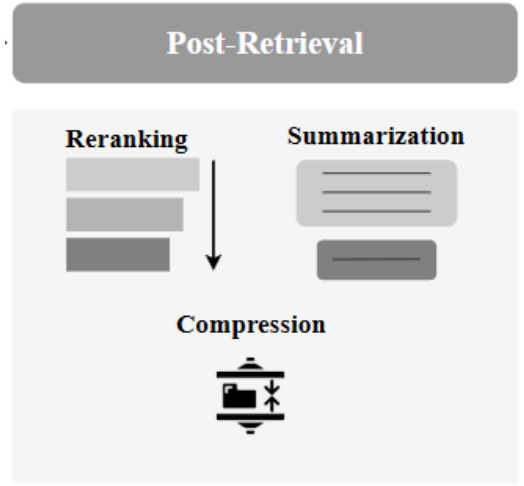
[1] Joko, Hideaki, et al. "Doing personal laps: Llm-augmented dialogue construction for personalized multi-session conversational search." Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024.

Post Retrieval

- Motivation
  - ***Current post-retrieval methods focus on general relevance and coherence improvement, but they lack adaptation to individual user preferences, limiting personalized experiences***

- Personalized Post-Retrieval Definition
  - Personalized post-retrieval integrates user-specific information to optimize the relevance, coherence, and user satisfaction of retrieved results

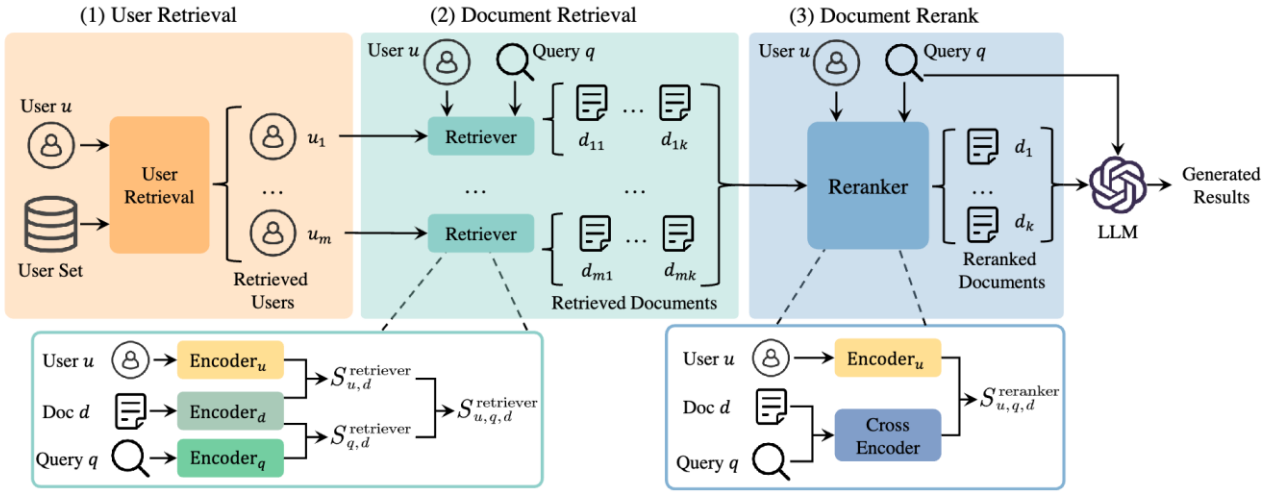
- Methods can be divided to:
  - Re-ranking
  - Summarization
  - Compression



Post Retrieval - Re-ranking

- CFRAG combines retrieval-augmented generation with collaborative filtering to produce more personalized and context-aware text.

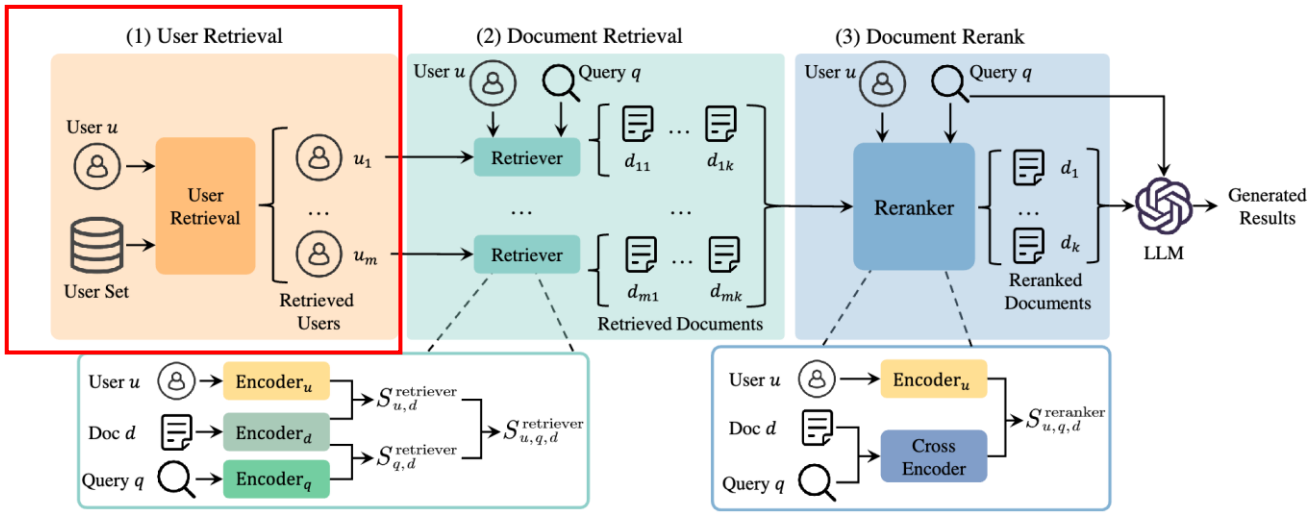
- CFRAG(SIGIR'25)
  - **Problem:**
    - Standard RAG re-ranks documents based on semantic relevance only — no awareness of who is asking
    - Retrieved candidates contain noise from a single user's history; static ranking ignores collaborative signals across similar users
  - **Method:**
    - Builds a three-stage pipeline
    - Personalized Reranker concatenates user embeddings with cross-encoder



[1] Shi, Teng, et al. "Retrieval augmented generation with collaborative filtering for personalized text generation." Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025

Post Retrieval - Re-ranking

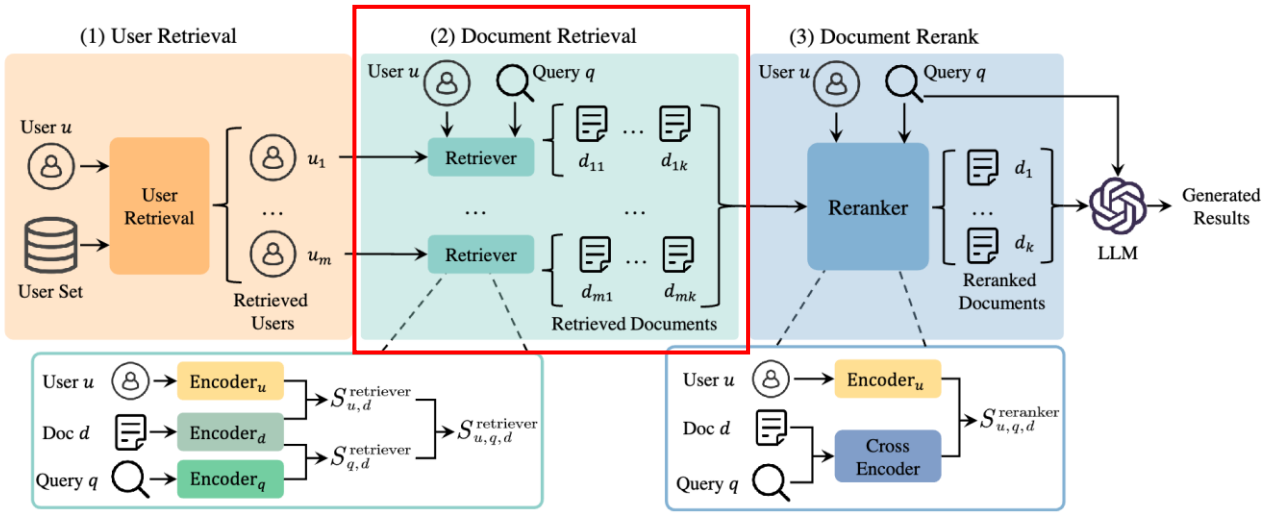
- Focusing on re-ranking retrieved documents based on live user profiles
- CFRAG(SIGIR'25)
  - **1. User Retrieval**
    - User history encoded via embedding model + Transformer to produce user embeddings
    - Three data augmentation strategies applied: Document Crop, Mask, and Reorder
    - Augmented views of the same user treated as positive pairs; other users as negatives



[1] Shi, Teng, et al. "Retrieval augmented generation with collaborative filtering for personalized text generation." Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025

Post Retrieval - Re-ranking

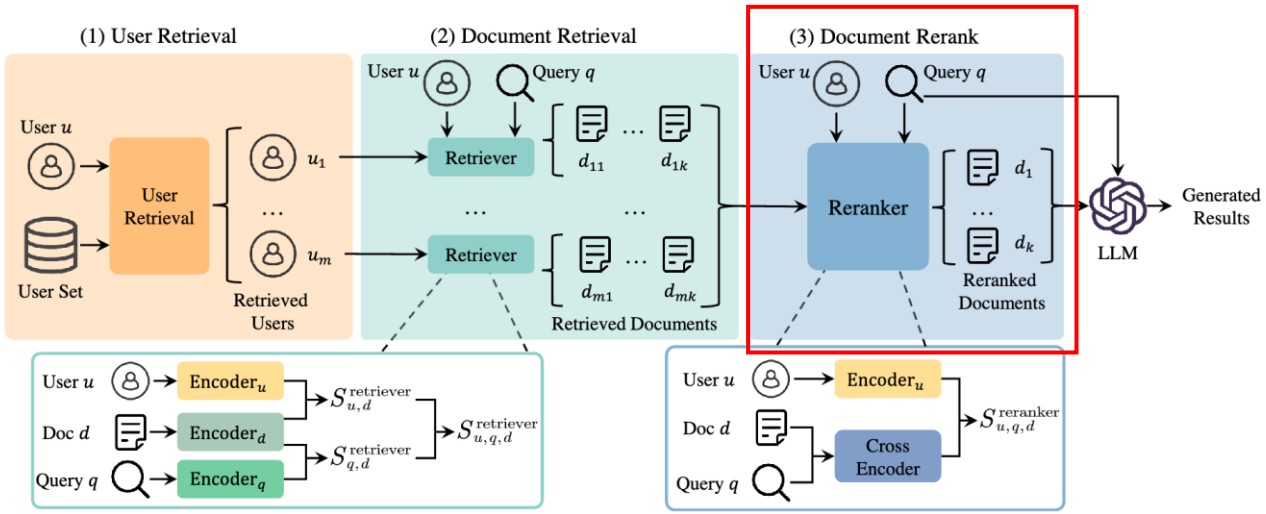
- Focusing on re-ranking retrieved documents based on live user profiles
- CFRAG(SIGIR'25)
  - **2. Personalized Document Retrieval**
    - Retrieves top-k documents from each of the m users' histories
    - Retrieval score combines semantic relevance (query–document) and user preference score (user–document)
    - User preference score maps user embedding into document encoding space via MLP



[1] Shi, Teng, et al. "Retrieval augmented generation with collaborative filtering for personalized text generation." Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025

Post Retrieval - Re-ranking

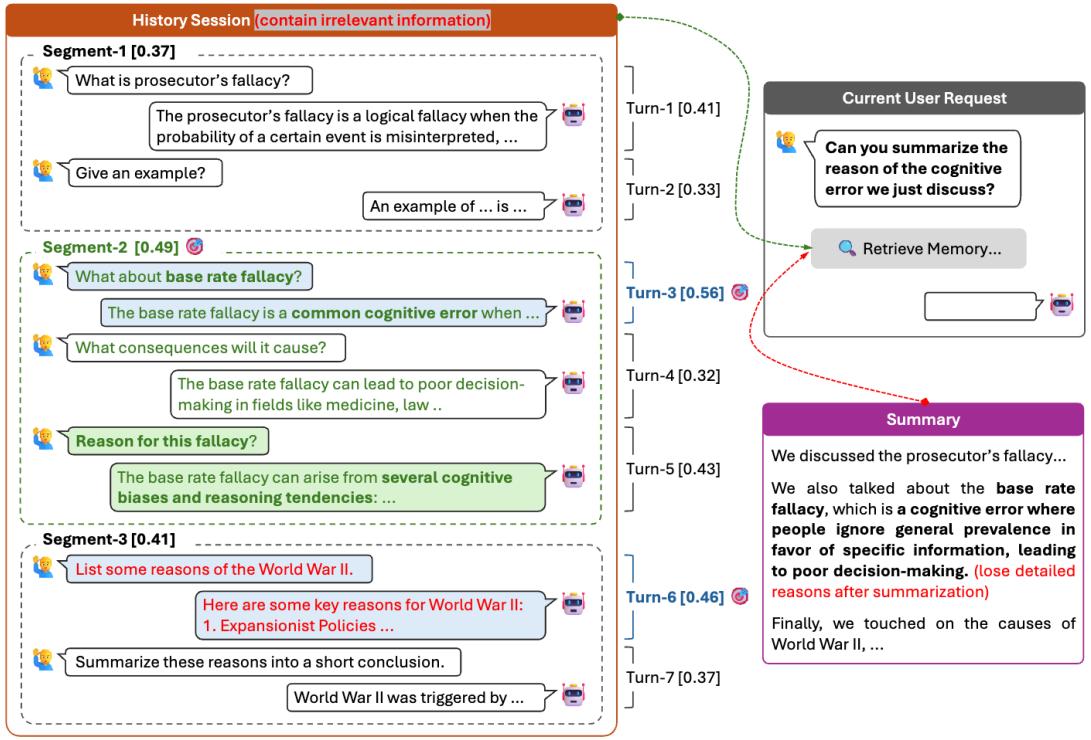
- Focusing on re-ranking retrieved documents based on live user profiles
- CFRAG(SIGIR'25)
  - **3. Personalized Document Reranking**
    - Cross-encoder encodes query–document pairs to compute semantic relevance
    - User embedding concatenated with cross-encoder output to inject personalized preference into ranking
    - Final reranking score produced by MLP over the combined representation



[1] Shi, Teng, et al. "Retrieval augmented generation with collaborative filtering for personalized text generation." Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2025

Post Retrieval - Compression

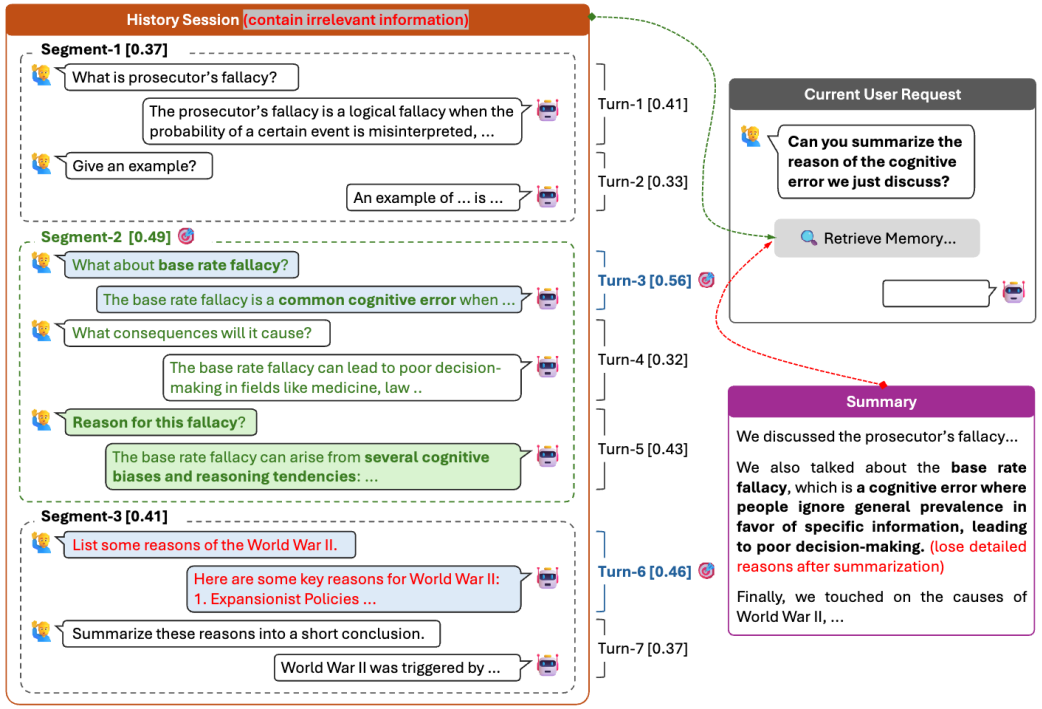
- SECOM enhances personalized memory retrieval through **compression-based denoising**, removing linguistic redundancy to improve retrieval accuracy and the quality of context ultimately fed to the LLM.
- SeCom (ICLR'25)
  - **Problem:**
    - conversational memories often contain a lot of linguistic redundancy
    - Even after memory segmentation, these noisy tokens may weaken the semantic signal and make retrieval less accurate
  - **Method:**
    - Introduces a conversation segmentation model
    - Applies LLMLingua-2 compression
    - Retrieved segment-level memory units concatenated directly as context



[1] Pan, Zhuoshi, et al. "SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents." The Thirteenth International Conference on Learning Representations.

Post Retrieval - Compression

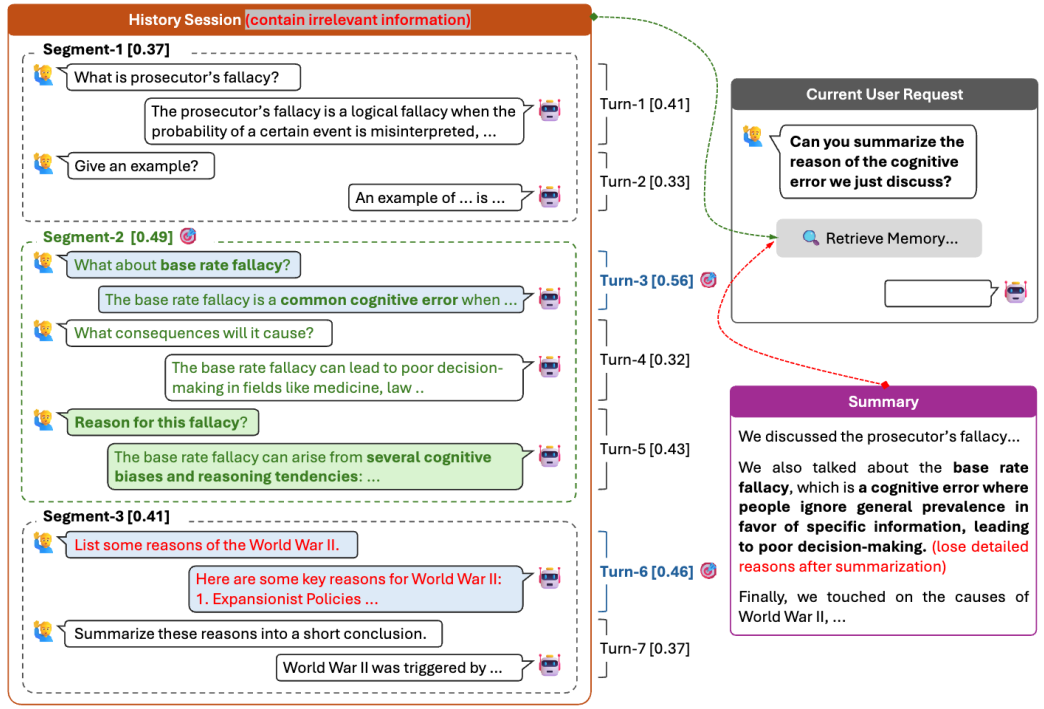
- SECOM enhances personalized memory retrieval through **compression-based denoising**, removing linguistic redundancy to improve retrieval accuracy and the quality of context ultimately fed to the LLM.
- SeCom (ICLR'25)
  - **1. Segment-Level Memory Construction**
    - Long-term conversation history partitioned into topically coherent segments as memory units
    - Each segment groups consecutive user-agent exchanges on the same topic
    - Segment boundaries identified when topic shifts occur



[1] Pan, Zhuoshi, et al. "SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents." The Thirteenth International Conference on Learning Representations.

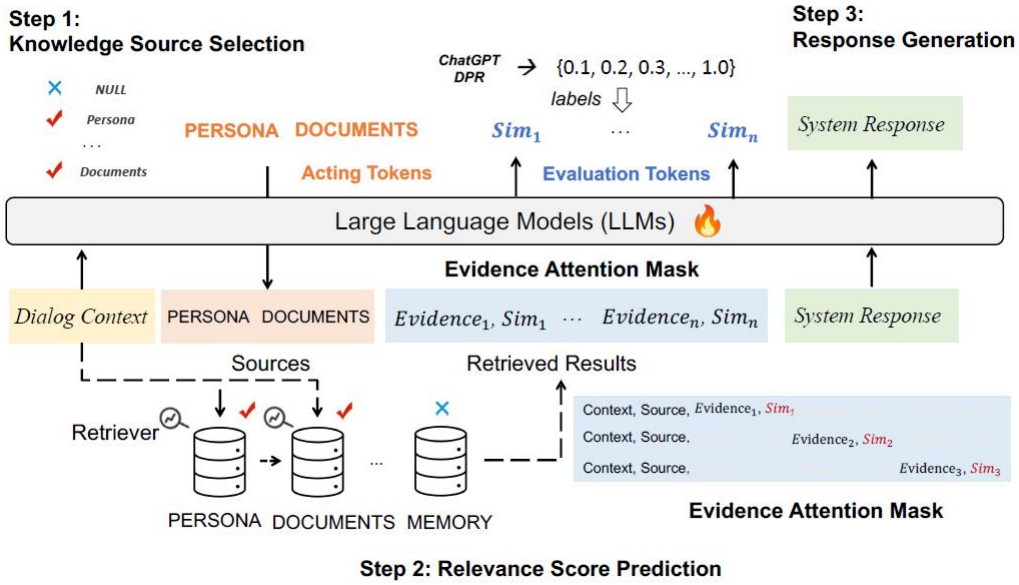
Post Retrieval - Compression

- SECOM enhances personalized memory retrieval through **compression-based denoising**, removing linguistic redundancy to improve retrieval accuracy and the quality of context ultimately fed to the LLM.
- SeCom (ICLR'25)
  - **2. Compression-Based Memory Denoising**
    - LLMLingua-2 applied to all memory units as a denoising function before retrieval
    - Compression rate set to 75% — removes linguistic redundancy from memory units
    - Post-compression: similarity between query and relevant segments increases; similarity with irrelevant segments decreases

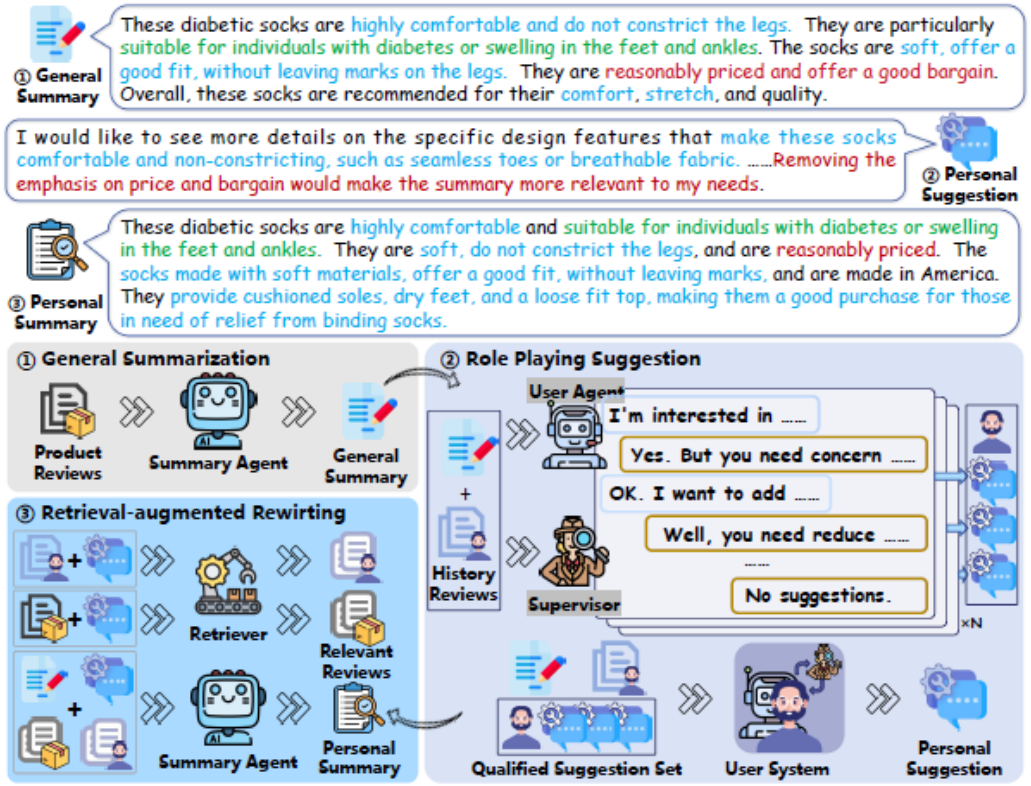


[1] Pan, Zhuoshi, et al. "SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents." The Thirteenth International Conference on Learning Representations.

Other Post Retrieval Works



➢ Overview of UniMS-RAG [1]



➢ Overview of PerChan [2]

[1] Wang, Hongru, et al. "Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems." arXiv preprint arXiv:2401.13256 (2024).
[2] Zhang, Yanyue, Yulan He, and Deyu Zhou. "Rehearse With User: Personalized Opinion Summarization via Role-Playing based on Large Language Models." arXiv preprint arXiv:2503.00449 (2025).

Overview

Method	Category	Method	Category
RECAP	Indexing	PAGE	Retrieval
EMG-RAG	Indexing	SAFARI	Retrieval
ROPG	Indexing	LAPS	Retrieval
PGraphRAG	Indexing	CFRAG	Post-Retrieval
PEARL	Indexing	SeCom	Post-Retrieval
RECAP	Indexing	UniMS-RAG	Post-Retrieval
XPERT	Retrieval	PerChan	Post-Retrieval
RF-Mem	Retrieval		

Retrieval Takeaways

- The key takeaways for retrieval are listed as follows:
  - Use dense retrieval methods like BERT and DeBERTa for high semantic understanding and personalization, or opt for prompt-based retrieval for flexible adaptation to user needs, though both require **significant computational resources** or **careful engineering**
  - Use reinforcement learning to dynamically adapt retrieval to user feedback and optimize results post-retrieval through re-ranking, summarization, or compression to **enhance personalization** and relevance while **balancing computational costs**

Agenda

1. Introduction
9:00 – 9:50



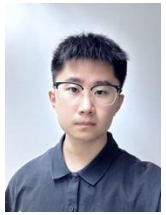
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



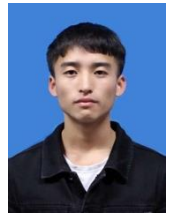
Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



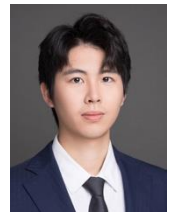
Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

8. Hands on Projects
16:45 – 17:30

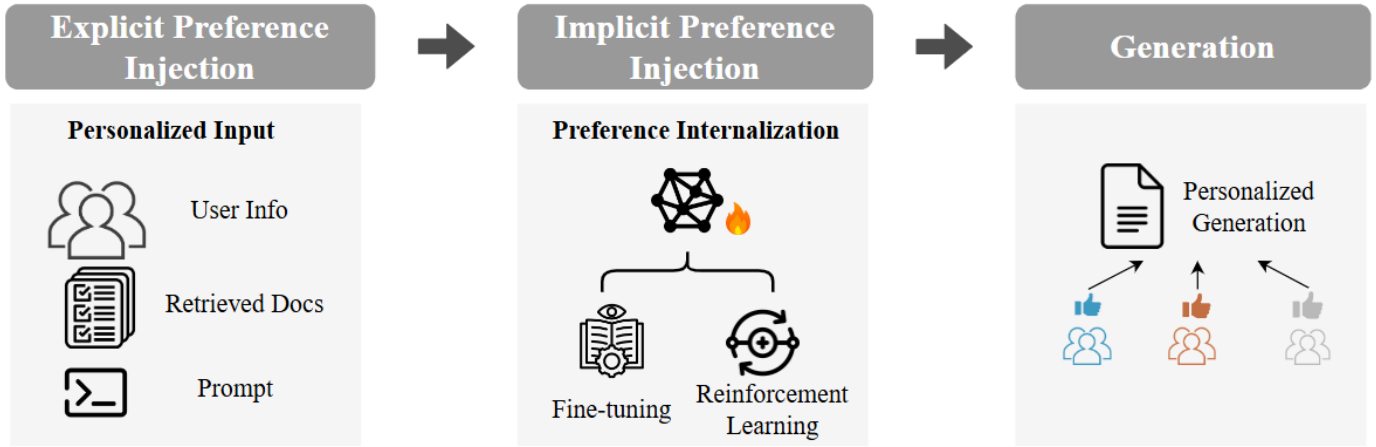


Pengyue Jia

Background

- Generation Definition:
  - Generation leverages ***personalized retrieved content, user information, and task context to generate responses*** that are better aligned with individual user needs and intent.

- Equation: $g^* = G(D^*, prompt, p, \theta)$
  - G is the generator
  - D^* is the personalized retrieved documents of the user
  - $prompt$ is the instruction for personalized generation
  - p is the personalization information of the user
  - θ is the parameter of generator
- According to Q, methods catrgories:
  - Generation from Explicit Preference
  - Generation from Implicit Preference



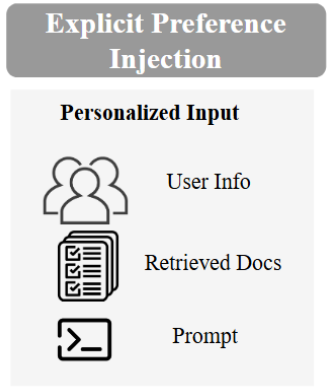
Personalization in Generation

- Generation Definition:
  - Generation leverages ***personalized retrieved content, user information, and task context to generate responses*** that are better aligned with individual user needs and intent.

- **Explicit Preference** injects user information, retrieved documents, and prompts as direct inputs to guide personalized generation.

$$g^* = \mathcal{G}\left(D^*, \text{prompt}, p, \theta\right)$$

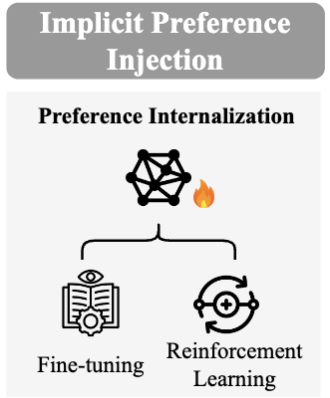
Explicit Preference



- **Implicit Preference** internalizes user preferences into model parameters through fine-tuning or reinforcement learning.

$$g^* = \mathcal{G}\left(D^*, \text{prompt}, p, \theta\right)$$

Implicit Preference



Generation from Explicit Preferences

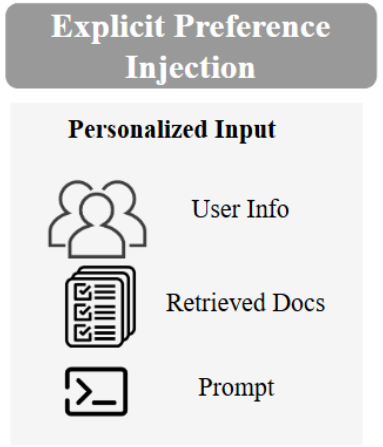
- Motivation
  - ***Traditional generation fail to integrate personalized user information .***

- Generation from Explicit Preferences
  - Generation from Explicit Preferences in personalized RAG refers to ***the integration of user personalized information*** p**, the retrieved personalized documents** D^***, and user-specific prompts** $prompt$ **into the generator to produce customized outputs.**

- Methods can be divided to:
  - Direct-Integrated Prompting
  - Summary-Augmented Prompting
  - Adaptive Prompting

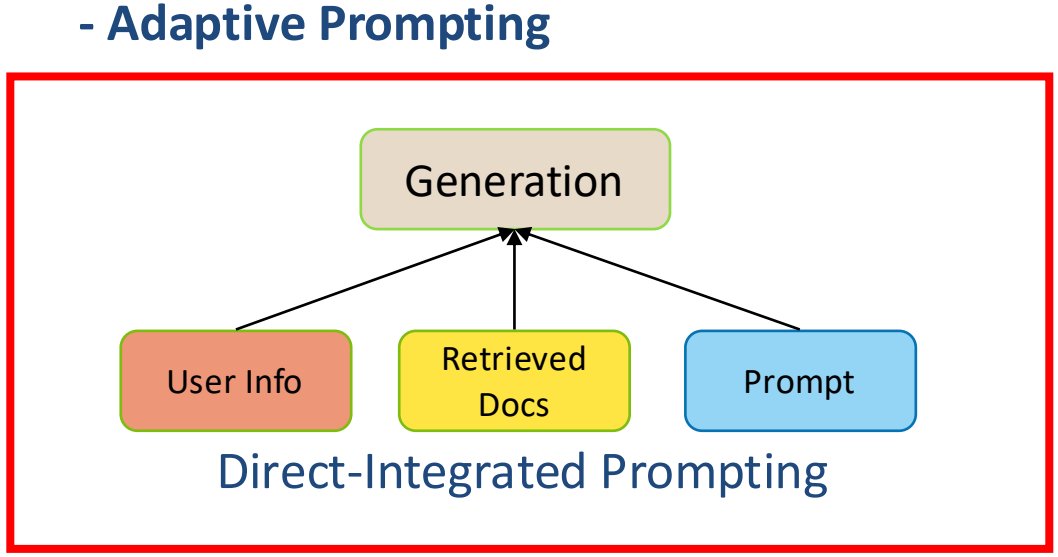
$$g^* = \mathcal{G}\left(D^*, \text{prompt}, p, \theta\right)$$

Explicit Preference

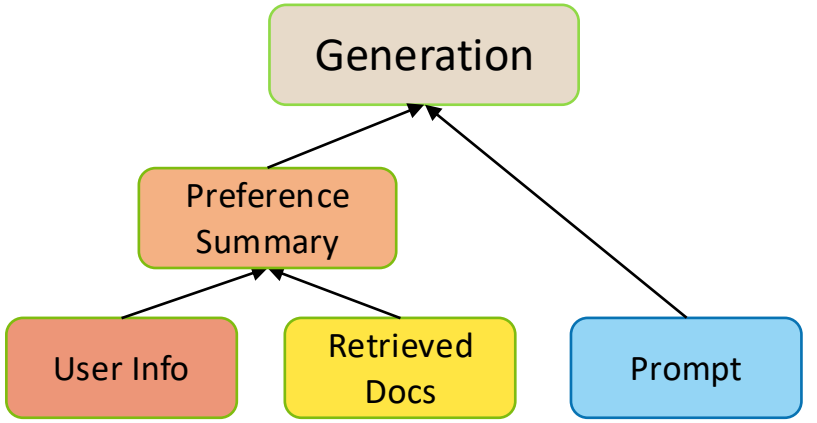


Generation from Explicit Preferences

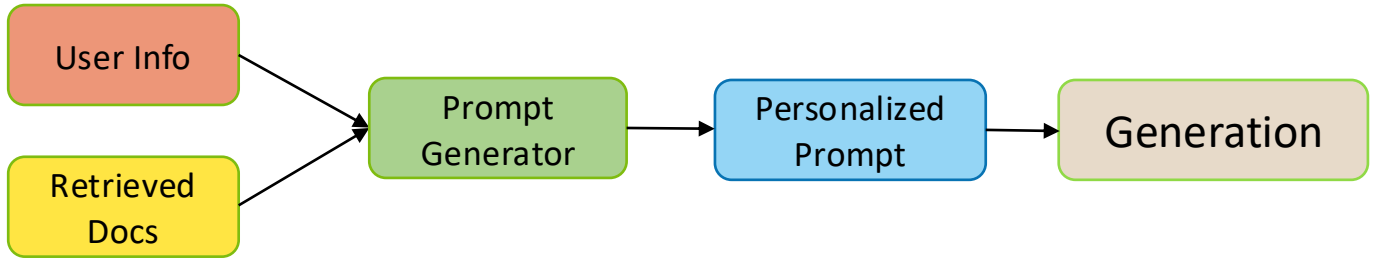
- Generation from Explicit Preference:
  - **Direct-Integrated Prompting**
  - **Summary-Augmented Prompting**
  - **Adaptive Prompting**



Direct-Integrated Prompting



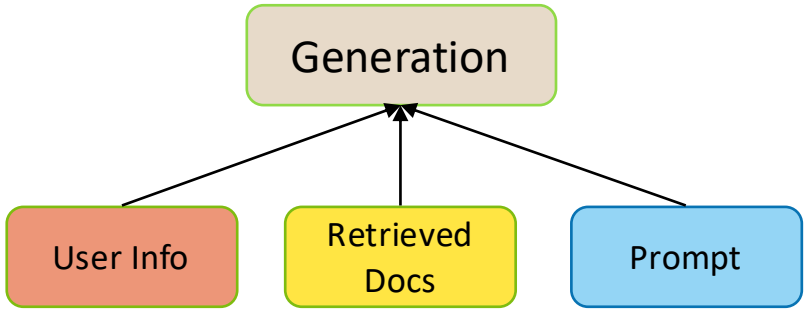
Summary-Augmented Prompting



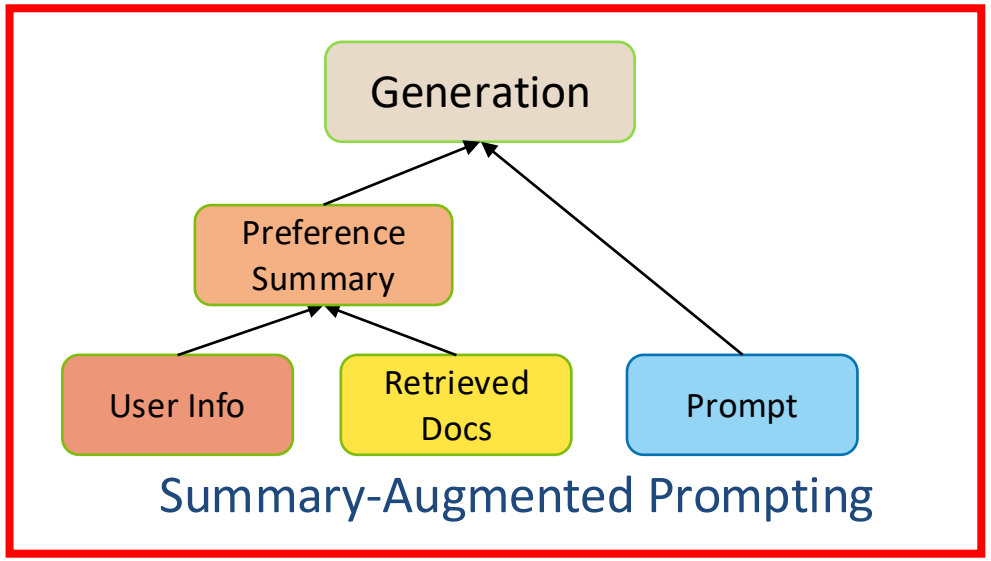
Adaptive Prompting

Generation from Explicit Preferences

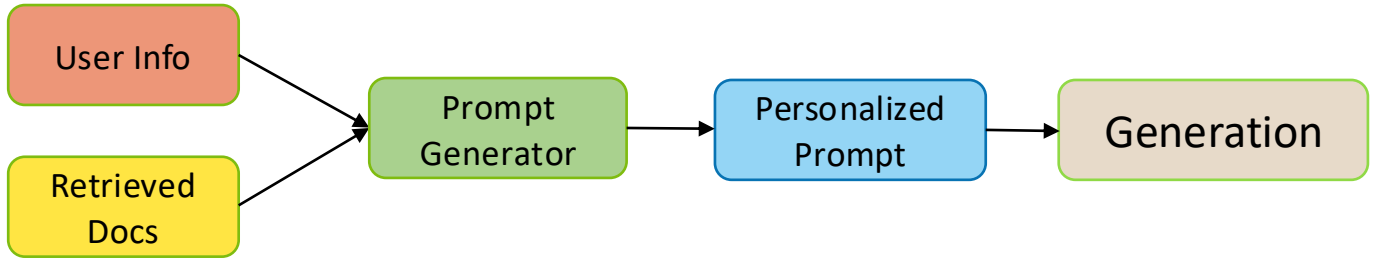
- Generation from Explicit Preference:
  - **Direct-Integrated Prompting**
  - **Summary-Augmented Prompting**
  - **Adaptive Prompting**



Direct-Integrated Prompting



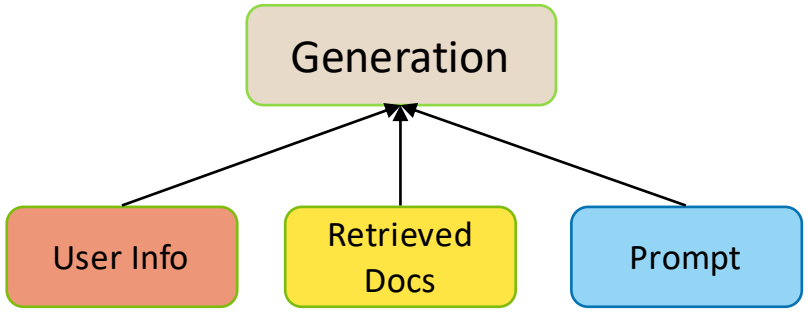
Summary-Augmented Prompting



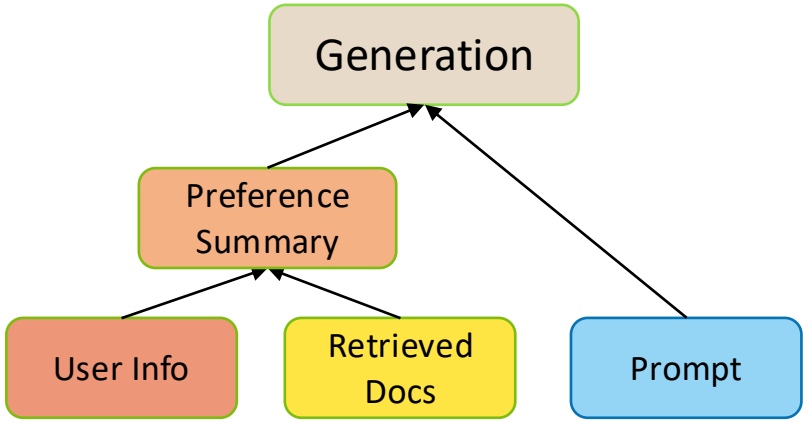
Adaptive Prompting

Generation from Explicit Preferences

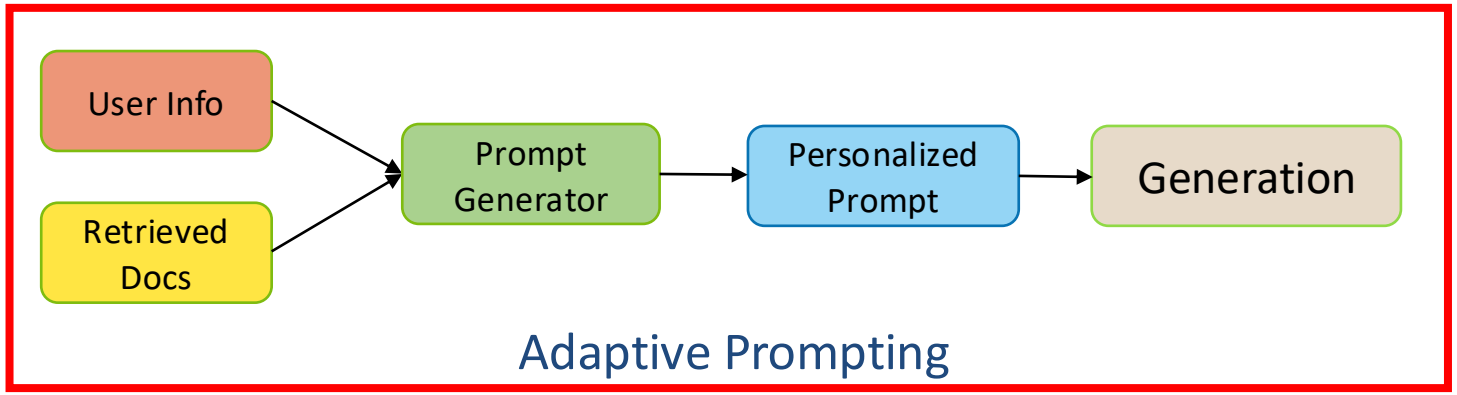
- Generation from Explicit Preference:
  - **Direct-Integrated Prompting**
  - **Summary-Augmented Prompting**
  - **Adaptive Prompting**



Direct-Integrated Prompting



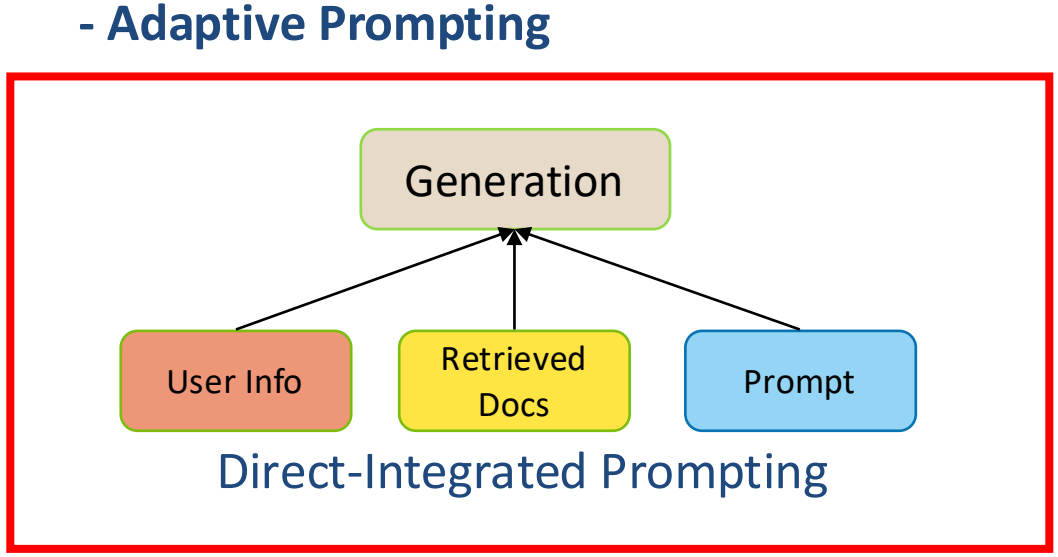
Summary-Augmented Prompting



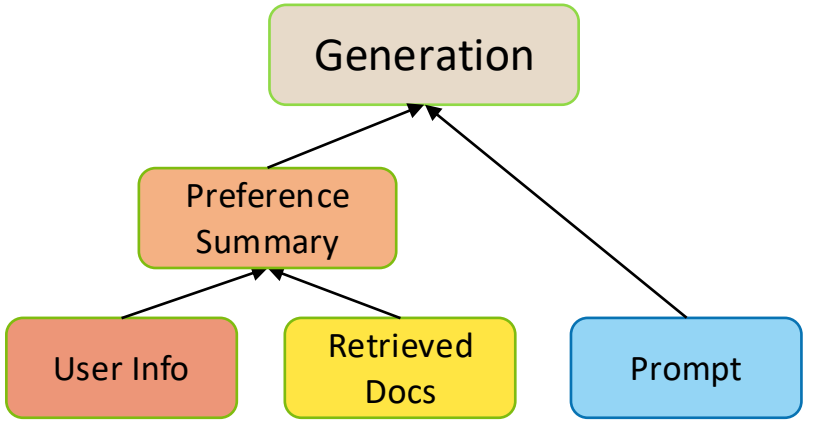
Adaptive Prompting

Generation from Explicit Preferences

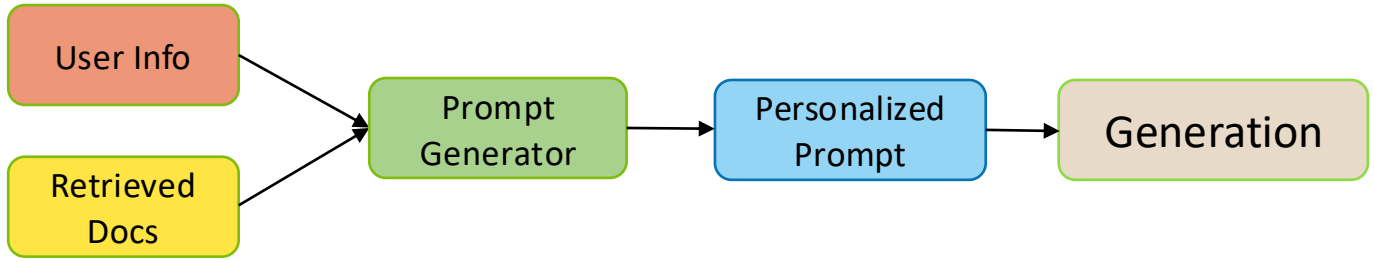
- Generation from Explicit Preference:
  - **Direct-Integrated Prompting**
  - **Summary-Augmented Prompting**
  - **Adaptive Prompting**



Direct-Integrated Prompting



Summary-Augmented Prompting



Adaptive Prompting

Cue-CoT

- **Motivation:**
  - Directly generating responses from dialogue context may miss the user's hidden status.
  - In-depth dialogue needs understanding user personality, emotion, and psychology.

- **Target:**
  - Dialogue Context → User Status Reasoning → Personalized Response

- **Methods:**
  - **O-Cue CoT:** output user status and response together.
  - **M-Cue CoT:** first infer user status, then generate response step-by-step.
  - **Cue-based Demonstration Selection:** select few-shot examples using dialogue context and inferred user status.

[1] Wang, Hongru, et al. "Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

Cue-CoT Details

- **O-Cue CoT**:
  - Prompt the models to generate the intermediate reasoning about user status and final results.

***O-Cue* CoT Prompting in Dialogue**

Model Input

Here is the conversation between user and system.

{DIALOGUE_CONTEXT}

Please first output a single line containing user status such as the user's personality traits, psychological and emotional states exhibited in the conversation. In the subsequent line, please play a role as system and generate a response based on the user status and the dialogue context.

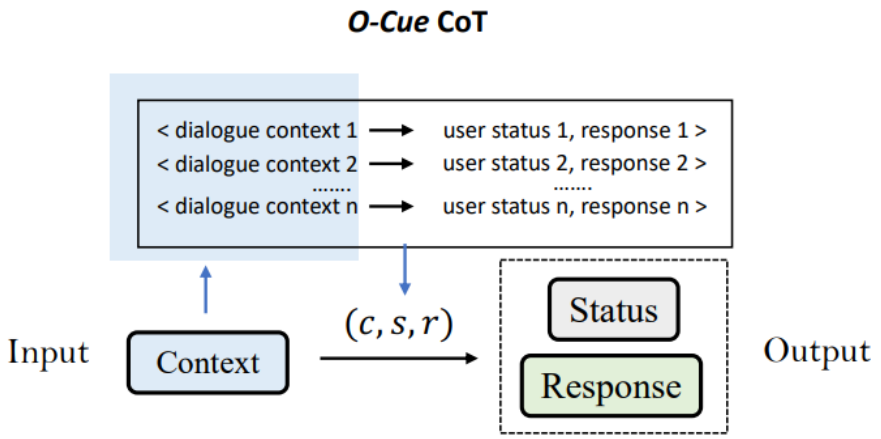
Model Output

User Status: *The status shown by the user is curiosity, a mindset of curiosity and eagerness to learn.*

Intermediate reasoning

System: *Sure! For learning the violin and piano from scratch, the most important thing is your learning attitude and patience in practice. Violin skills are relatively high, and it may be difficult to master them from scratch, but as long as you maintain a conscious learning spirit and continuous practice, you can gradually master violin skills. For piano, it is relatively easier to learn. As long as you have a certain sense of rhythm and music foundation, it is feasible to learn piano from scratch.*

SCORES 8.5 / 10



[1] Wang, Hongru, et al. "Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

Cue-CoT Details

- **O-Cue CoT**:
  - Prompt the models to generate the intermediate reasoning about user status and final results.
- **Demonstration Selection:**
  - Choose the most contextually relevant sample (c, s, r) based on current context.

***O-Cue* CoT Prompting in Dialogue**

Model Input

Here is the conversation between user and system.

{DIALOGUE_CONTEXT}

Please first output a single line containing user status such as the user's personality traits, psychological and emotional states exhibited in the conversation. In the subsequent line, please play a role as system and generate a response based on the user status and the dialogue context.

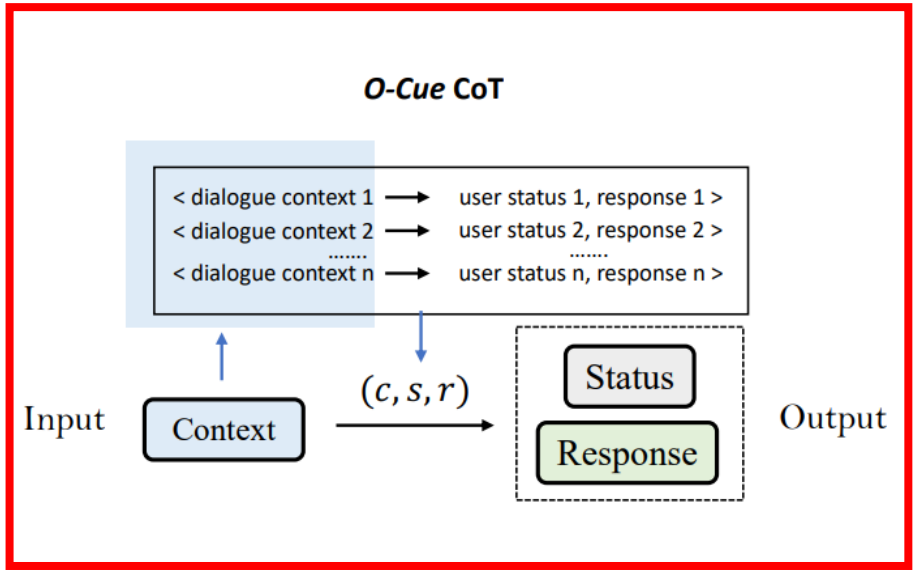
Model Output

User Status: *The status shown by the user is curiosity, a mindset of curiosity and eagerness to learn.*

Intermediate reasoning

System: *Sure! For learning the violin and piano from scratch, the most important thing is your learning attitude and patience in practice. Violin skills are relatively high, and it may be difficult to master them from scratch, but as long as you maintain a conscious learning spirit and continuous practice, you can gradually master violin skills. For piano, it is relatively easier to learn. As long as you have a certain sense of rhythm and music foundation, it is feasible to learn piano from scratch.*

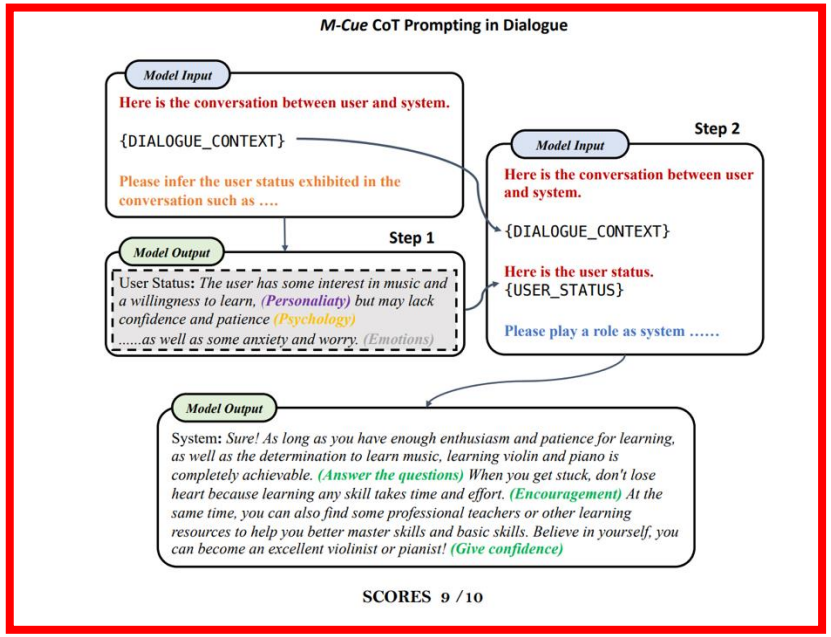
SCORES 8.5 / 10

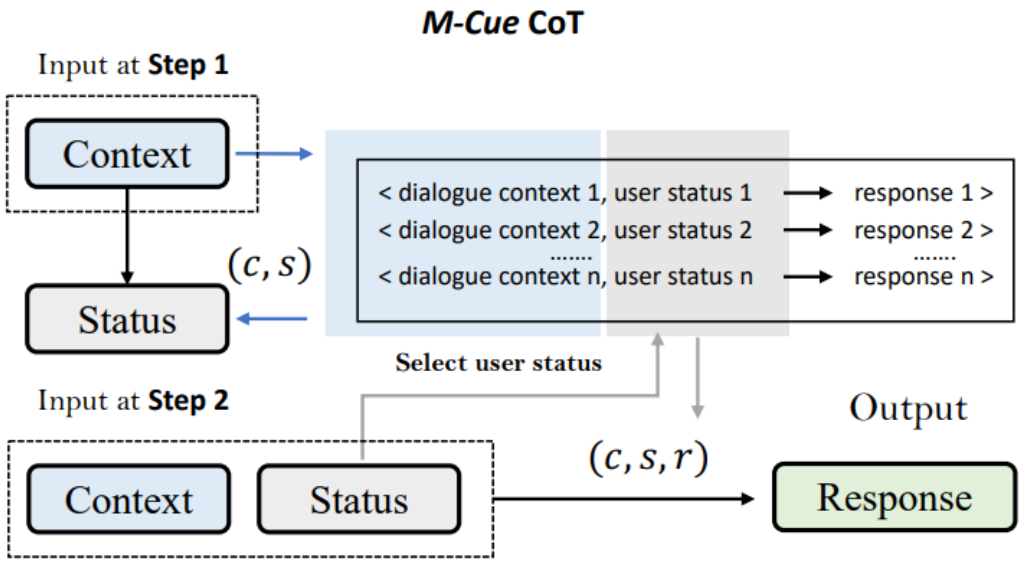


[1] Wang, Hongru, et al. "Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

Cue-CoT Details

- **M-Cue CoT**:
  - Step 1: Infer user status based on dialogue context.
  - Step 2: Output response based on the context and generated user status.

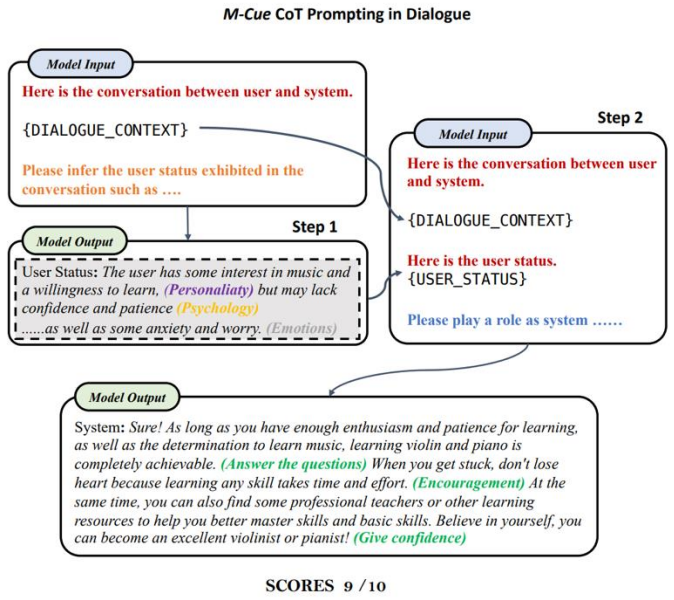


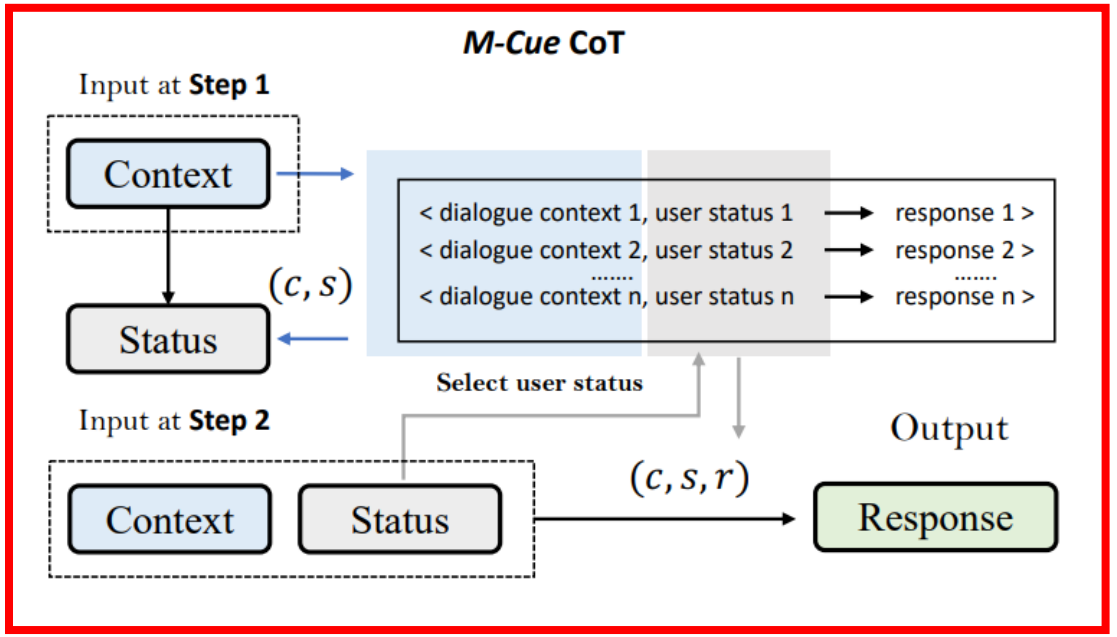


[1] Wang, Hongru, et al. "Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

Cue-CoT Details

- **M-Cue CoT**:
  - Step 1: Infer user status based on dialogue context.
  - Step 2: Output response based on the context and generated user status.
- **Demonstration Selection:**
  - For step 1: Retrieve context-user status pair (c, s) based on current context.
  - For step 2: Retrieve (c, s, r) based on current context and user status.





[1] Wang, Hongru, et al. "Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs." Findings of the Association for Computational Linguistics: EMNLP 2023. 2023.

Cue-CoT Findings

- **Cue-CoT improves over standard prompting**
  - Both O-Cue and M-cue methods outperform standatd prompting, which proofing the effectiveness of explicit user-status reasoning.

Model	Prompt	Helpfulness			Acceptability		
		Quora	ED	EMH	Quora	ED	EMH
Zero-shot Setting							
ALPACA	*O-Cue*	19.51	39.41	49.70	22.85	35.41	50.15
	M-Cue	80.78	87.30	85.76	78.21	86.00	86.97
VICUNA	*O-Cue*	56.16	71.43	59.43	55.73	65.06	63.50
	M-Cue	81.67	91.30	80.42	77.89	90.71	82.93
CHATGPT	*O-Cue*	79.47	88.31	82.83	81.47	89.92	93.71
	M-Cue	85.83	91.98	82.93	89.09	96.79	94.93
One-shot Setting							
							random selection
ALPACA	*O-Cue*	-	-	-	-	-	-
	M-Cue	76.78	85.08	94.36	72.34	85.07	95.82
VICUNA	*O-Cue*	60.45	70.77	63.06	60.45	68.21	67.07
	M-Cue	79.84	91.20	79.23	83.16	92.45	87.99
CHATGPT	*O-Cue*	80.33	87.32	84.94	80.33	90.80	96.06
	M-Cue	84.31	89.78	85.71	86.64	93.94	96.70
							top-1 selection
ALPACA	*O-Cue*	-	-	-	-	-	-
	M-Cue	74.54	78.70	88.69	72.27	79.55	93.43
VICUNA	*O-Cue*	63.10	71.75	62.31	62.04	67.21	67.76
	M-Cue	78.70	90.12	79.10	82.08	92.96	88.96
CHATGPT	*O-Cue*	81.15	87.42	81.40	80.24	89.92	91.84
	M-Cue	88.08	91.37	86.87	91.21	95.95	96.12

Report the win rate of Cue-Cot compared with standard prompting

Cue-CoT Findings

- **Cue-CoT improves over standard prompting**
  - Both O-Cue and M-cue methods outperform standatd prompting, which proofing the effectiveness of explicit user-status reasoning.

- **M-Cue is better than O-Cue**
  - It decomposes the task into “infer user status → generate response,” making the status an explicit condition.

Model	Prompt	Helpfulness			Acceptability		
		Quora	ED	EMH	Quora	ED	EMH
Zero-shot Setting							
ALPACA	*O-Cue*	19.51	39.41	49.70	22.85	35.41	50.15
	M-Cue	80.78	87.30	85.76	78.21	86.00	86.97
VICUNA	*O-Cue*	56.16	71.43	59.43	55.73	65.06	63.50
	M-Cue	81.67	91.30	80.42	77.89	90.71	82.93
CHATGPT	*O-Cue*	79.47	88.31	82.83	81.47	89.92	93.71
	M-Cue	85.83	91.98	82.93	89.09	96.79	94.93
One-shot Setting							
							random selection
ALPACA	*O-Cue*	-	-	-	-	-	-
	M-Cue	76.78	85.08	94.36	72.34	85.07	95.82
VICUNA	*O-Cue*	60.45	70.77	63.06	60.45	68.21	67.07
	M-Cue	79.84	91.20	79.23	83.16	92.45	87.99
CHATGPT	*O-Cue*	80.33	87.32	84.94	80.33	90.80	96.06
	M-Cue	84.31	89.78	85.71	86.64	93.94	96.70
							top-1 selection
ALPACA	*O-Cue*	-	-	-	-	-	-
	M-Cue	74.54	78.70	88.69	72.27	79.55	93.43
VICUNA	*O-Cue*	63.10	71.75	62.31	62.04	67.21	67.76
	M-Cue	78.70	90.12	79.10	82.08	92.96	88.96
CHATGPT	*O-Cue*	81.15	87.42	81.40	80.24	89.92	91.84
	M-Cue	88.08	91.37	86.87	91.21	95.95	96.12

Report the win rate of Cue-Cot compared with standard prompting

Cue-CoT Findings

- **Cue-CoT improves over standard prompting**
  - Both O-Cue and M-cue methods outperform standatd prompting, which proofing the effectiveness of explicit user-status reasoning.

- **M-Cue is better than O-Cue**
  - It decomposes the task into “infer user status → generate response,” making the status an explicit condition.

- **One-shot setting helps**
  - Examples help the model learn response format and style, especially when selected by similar user status

Model	Prompt	Helpfulness			Acceptability		
		Quora	ED	EMH	Quora	ED	EMH
Zero-shot Setting							
ALPACA	*O-Cue*	19.51	39.41	49.70	22.85	35.41	50.15
	M-Cue	80.78	87.30	85.76	78.21	86.00	86.97
VICUNA	*O-Cue*	56.16	71.43	59.43	55.73	65.06	63.50
	M-Cue	81.67	91.30	80.42	77.89	90.71	82.93
CHATGPT	*O-Cue*	79.47	88.31	82.83	81.47	89.92	93.71
	M-Cue	85.83	91.98	82.93	89.09	96.79	94.93
One-shot Setting							
random selection							
ALPACA	*O-Cue*	-	-	-	-	-	-
	M-Cue	76.78	85.08	94.36	72.34	85.07	95.82
VICUNA	*O-Cue*	60.45	70.77	63.06	60.45	68.21	67.07
	M-Cue	79.84	91.20	79.23	83.16	92.45	87.99
CHATGPT	*O-Cue*	80.33	87.32	84.94	80.33	90.80	96.06
	M-Cue	84.31	89.78	85.71	86.64	93.94	96.70
top-1 selection							
ALPACA	*O-Cue*	-	-	-	-	-	-
	M-Cue	74.54	78.70	88.69	72.27	79.55	93.43
VICUNA	*O-Cue*	63.10	71.75	62.31	62.04	67.21	67.76
	M-Cue	78.70	90.12	79.10	82.08	92.96	88.96
CHATGPT	*O-Cue*	81.15	87.42	81.40	80.24	89.92	91.84
	M-Cue	88.08	91.37	86.87	91.21	95.95	96.12

Report the win rate of Cue-Cot compared with standard prompting

TICL(NAACL'25)

- **Motivation:**
  - LLMs are aligned to generic preferences and often miss individual writing styles.
  - TEFT-style personalization requires per-user SFT/DPO updates, making it costly and impractical.

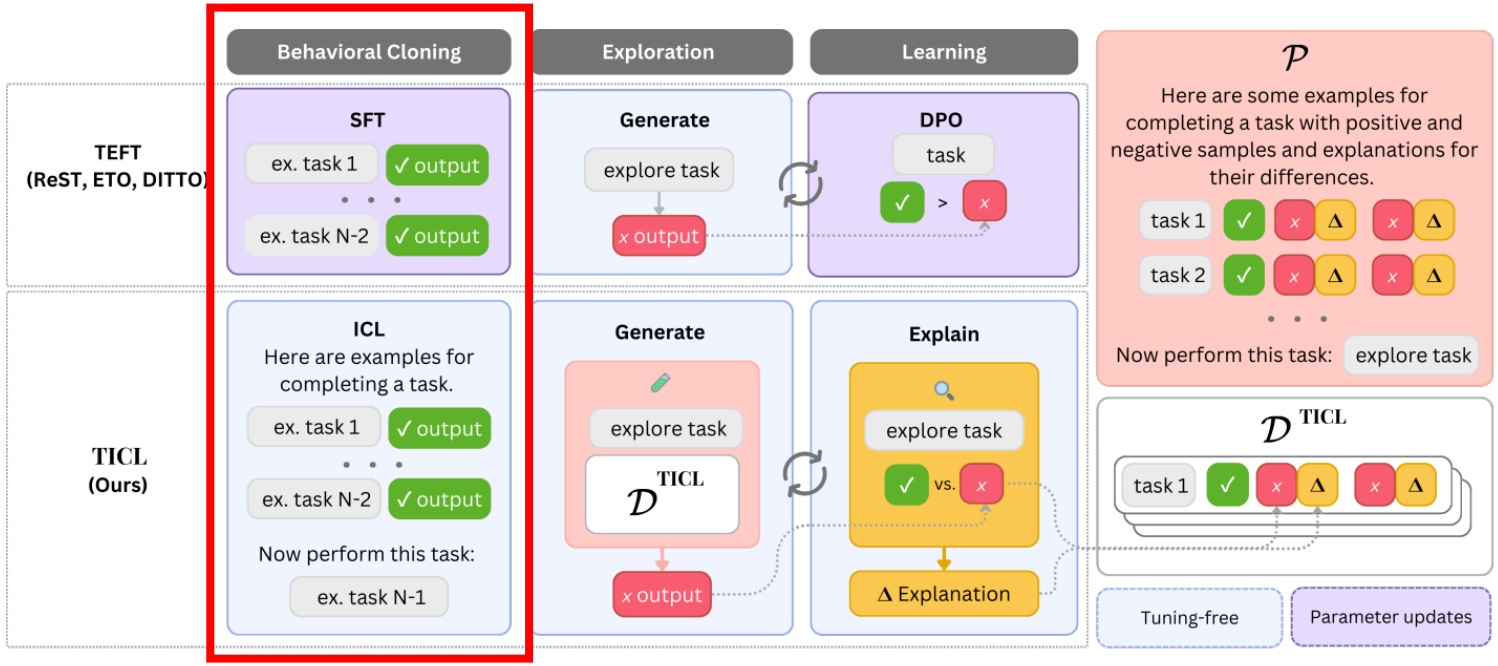
- **Target:**
  - Achieve tuning-free personalized alignment with fewer than 10 user examples, using prompt-level trial-and-error learning instead of parameter updates.

- **Methods:**
  - **Behavioral Cloning** uses positive user examples as few-shot ICL;
  - **Exploration** generates trial outputs to expose style mismatches;
  - **Learning** adds negative outputs and error explanations back into the prompt to teach what to avoid.

[1] Cho, Hyundong Justin, et al. "Tuning-free personalized alignment via trial-error-explain in-context learning." Findings of the Association for Computational Linguistics: NAACL 2025. 2025.

TICL Details

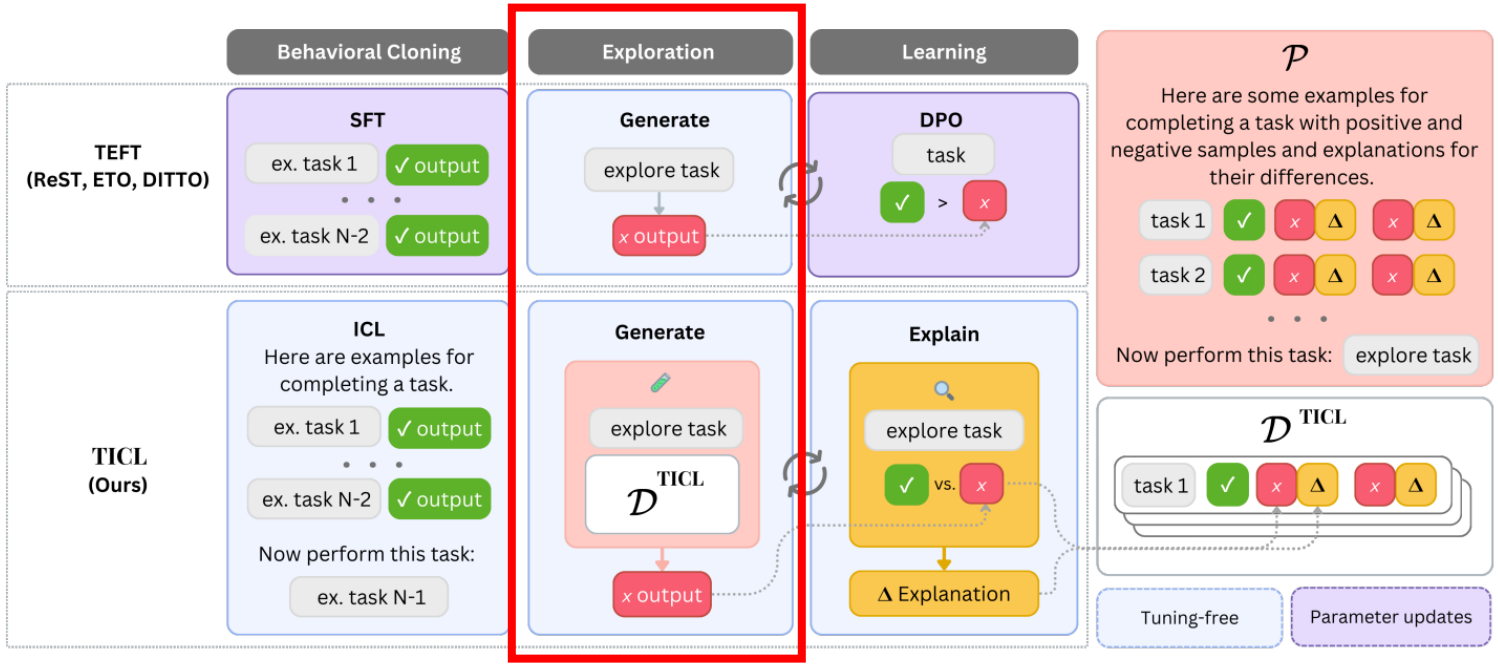
- **Behavioral Cloning / Style ICL**:
  - Imitate the user's writing style from positive few-shot examples.



[1] Cho, Hyundong Justin, et al. "Tuning-free personalized alignment via trial-error-explain in-context learning." Findings of the Association for Computational Linguistics: NAACL 2025. 2025.

TICL Details

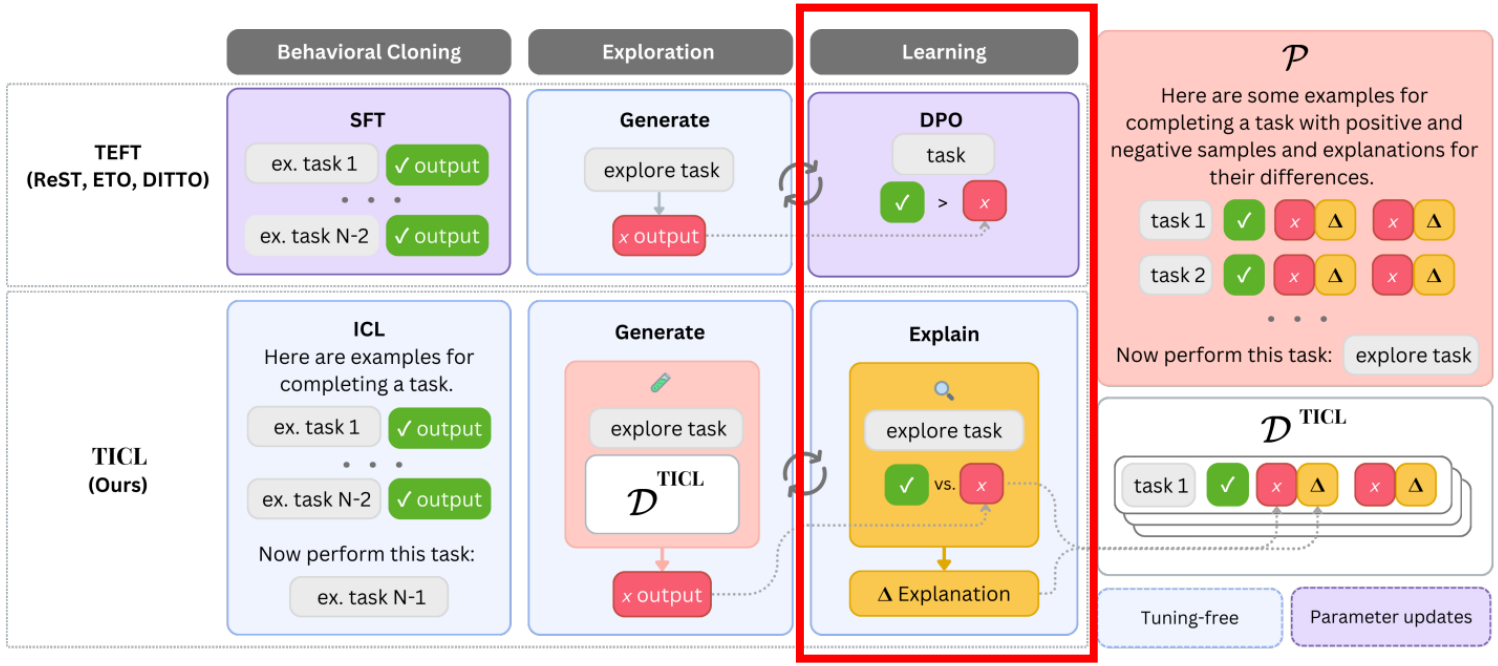
- **Behavioral Cloning / Style ICL**:
  - Imitate the user's writing style from positive few-shot examples.
- **Exploration / Trial Generation:**
  - Creating a candidate negative sample to expose style mismatches.



[1] Cho, Hyundong Justin, et al. "Tuning-free personalized alignment via trial-error-explain in-context learning." Findings of the Association for Computational Linguistics: NAACL 2025. 2025.

TICL Details

- **Behavioral Cloning / Style ICL**:
  - Imitate the user's writing style from positive few-shot examples.
- **Exploration / Trial Generation:**
  - Creating a candidate negative sample to expose style mismatches.
- **Learning / Error-Explanation Update**
  - Add failed outputs and explanations as feedback demonstrations.



[1] Cho, Hyundong Justin, et al. "Tuning-free personalized alignment via trial-error-explain in-context learning." Findings of the Association for Computational Linguistics: NAACL 2025. 2025.

TICL Findings

- **Error Explanations are the key signal**
  - Removing explanations drops GPT-4o 31.0→23.5 and Claude 54.5→46.0.

Model	Ablation	Win rate
GPT-4o	TICL	$\mathbf{31.00_{1.18}}$
	– Initial ICL examples	$28.50_{1.50}$
	– Explanations	$23.50_{1.24}$
	– Checkpointing	$22.50_{0.90}$
	– Negative samples & Expl.	$21.00_{1.28}$
Claude 3 S	TICL	$\mathbf{54.50_{1.67}}$
	– Initial ICL examples	$52.00_{1.22}$
	– Explanations	$46.00_{1.41}$
	– Checkpointing	$54.00_{1.67}$
	– Negative samples & Expl.	$43.50_{1.37}$

Ablation Study

[1] Cho, Hyundong Justin, et al. "Tuning-free personalized alignment via trial-error-explain in-context learning." Findings of the Association for Computational Linguistics: NAACL 2025. 2025.

TICL Findings

- **Error Explanations are the key signal**
  - Removing explanations drops GPT-4o 31.0→23.5 and Claude 54.5→46.0.
- **Negative feedback beats pure few-shot**
  - Removing negative samples + explanations drops to 21.0 / 43.5, close to plain few-shot.

Model	Ablation	Win rate
GPT-4o	TICL	$\mathbf{31.00_{1.18}}$
	– Initial ICL examples	$28.50_{1.50}$
	– Explanations	$23.50_{1.24}$
	– Checkpointing	$22.50_{0.90}$
	– Negative samples & Expl.	$21.00_{1.28}$
Claude 3 S	TICL	$\mathbf{54.50_{1.67}}$
	– Initial ICL examples	$52.00_{1.22}$
	– Explanations	$46.00_{1.41}$
	– Checkpointing	$54.00_{1.67}$
	– Negative samples & Expl.	$43.50_{1.37}$

Ablation Study

[1] Cho, Hyundong Justin, et al. "Tuning-free personalized alignment via trial-error-explain in-context learning." Findings of the Association for Computational Linguistics: NAACL 2025. 2025.

TICL Findings

- **Error Explanations are the key signal**
  - Removing explanations drops GPT-4o 31.0→23.5 and Claude 54.5→46.0.
- **Negative feedback beats pure few-shot**
  - Removing negative samples + explanations drops to 21.0 / 43.5, close to plain few-shot.
- **Few-shot ICL is only a starting point, not the key improvement**
  - Removing initial ICL examples causes small drops.

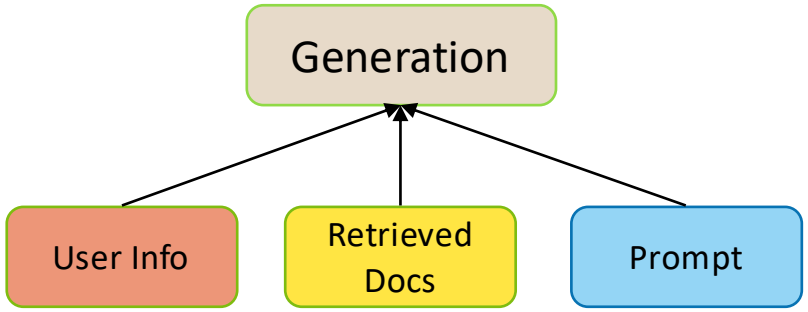
Model	Ablation	Win rate
GPT-4o	TICL	$\mathbf{31.00_{1.18}}$
	– Initial ICL examples	$28.50_{1.50}$
	– Explanations	$23.50_{1.24}$
	– Checkpointing	$22.50_{0.90}$
	– Negative samples & Expl.	$21.00_{1.28}$
Claude 3 S	TICL	$\mathbf{54.50_{1.67}}$
	– Initial ICL examples	$52.00_{1.22}$
	– Explanations	$46.00_{1.41}$
	– Checkpointing	$54.00_{1.67}$
	– Negative samples & Expl.	$43.50_{1.37}$

Ablation Study

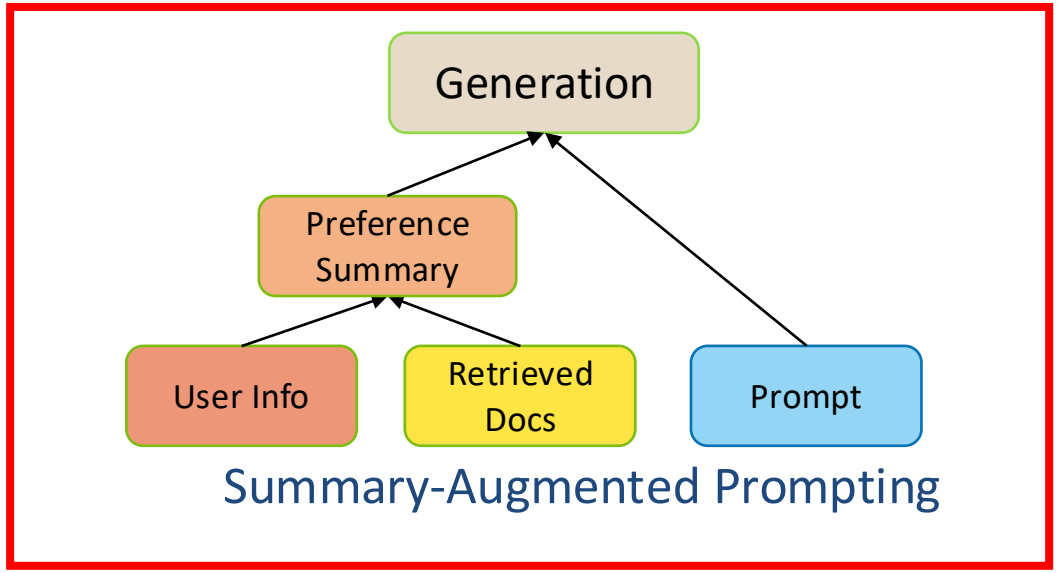
[1] Cho, Hyundong Justin, et al. "Tuning-free personalized alignment via trial-error-explain in-context learning." Findings of the Association for Computational Linguistics: NAACL 2025. 2025.

Generation from Explicit Preferences

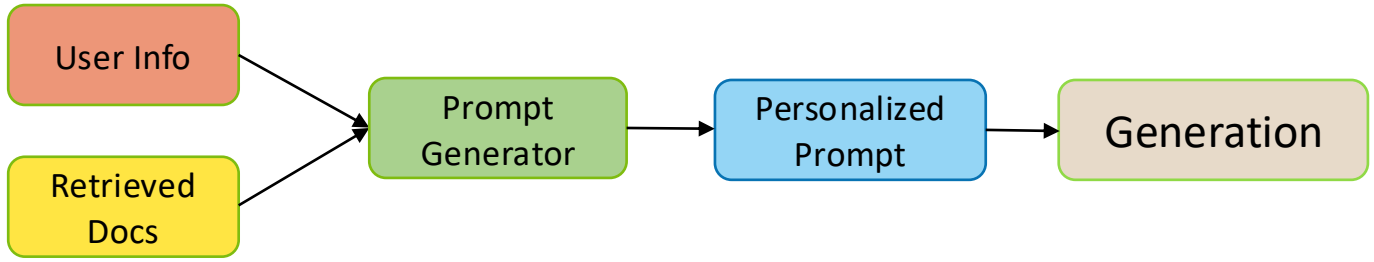
- Generation from Explicit Preference:
  - **Direct-Integrated Prompting**
  - **Summary-Augmented Prompting**
  - **Adaptive Prompting**



Direct-Integrated Prompting



Summary-Augmented Prompting



Adaptive Prompting

GPG

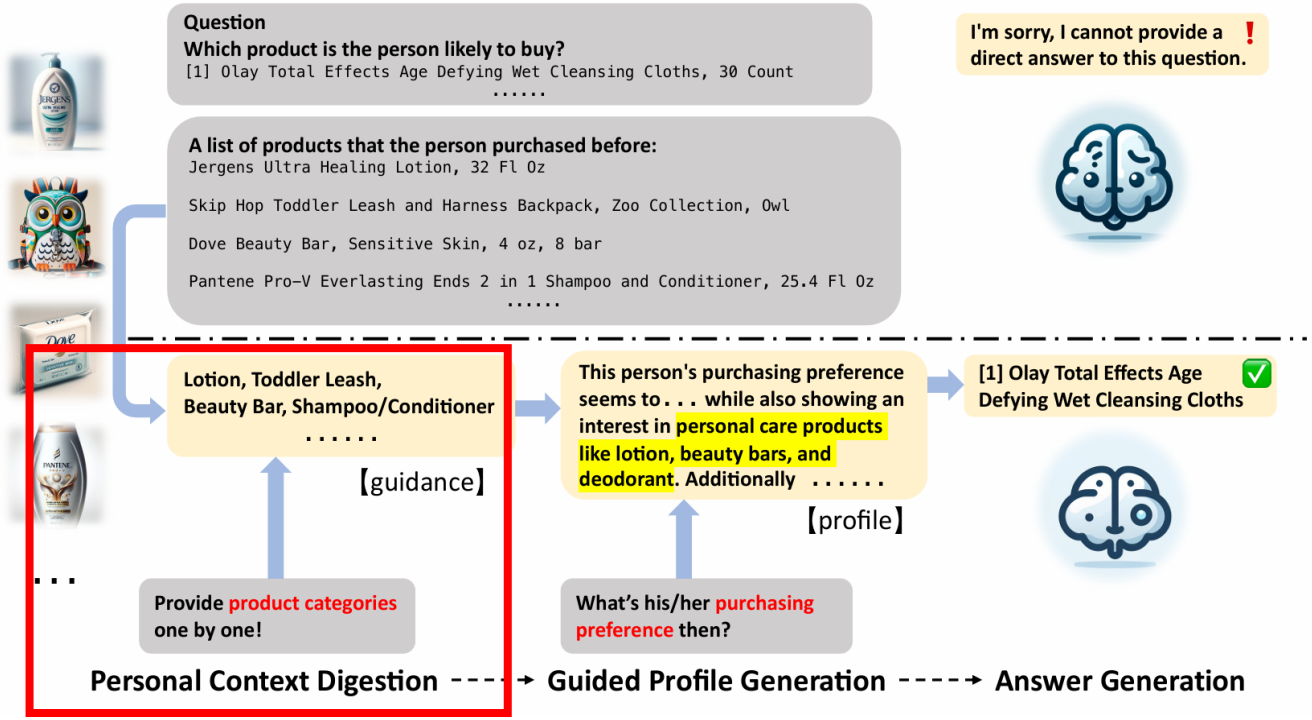
- **Motivation:**
  - LLMs struggle to understand sparse and noisy personal context.
  - Raw user history is long, costly, and hard to reuse.

- **Target:**
  - Compress user history into a concise personal profile for generation.

[1] Zhang, Jiarui. "Guided profile generation improves personalization with large language models." Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

GPG Details

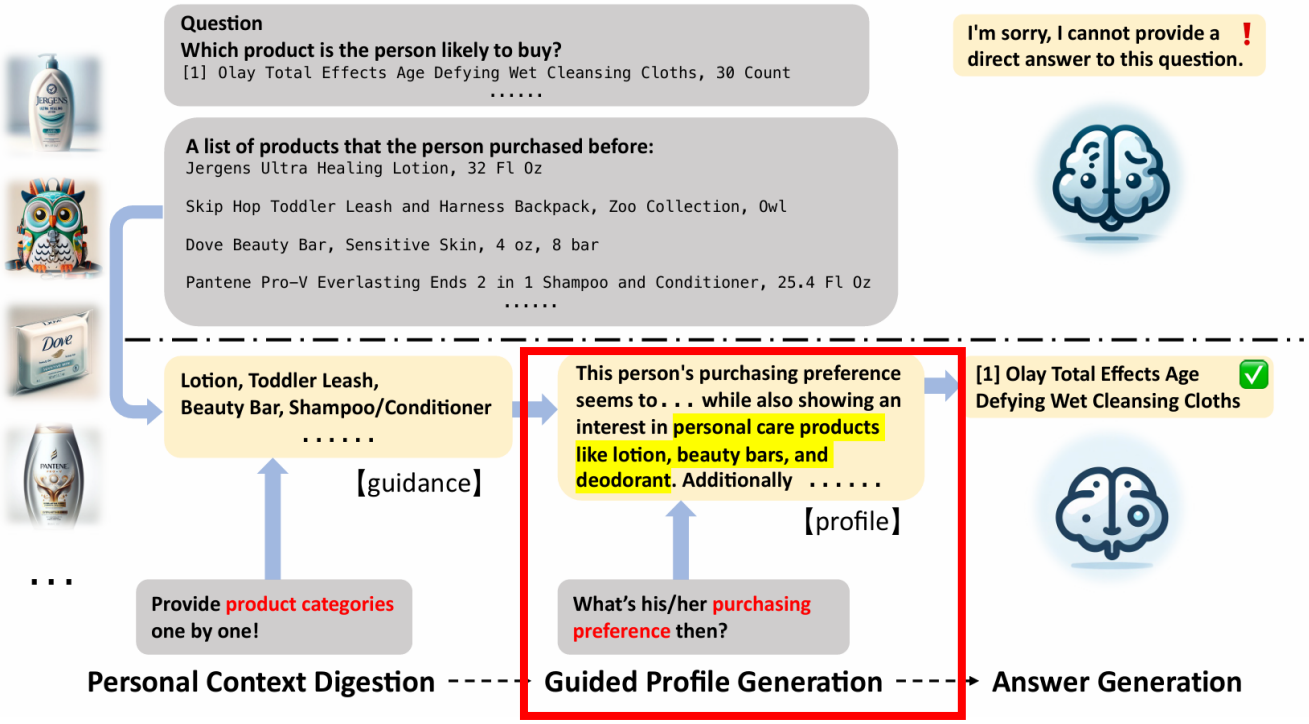
- **Personal Context Digestion:**
  - Extract useful user signals with task-specific instructions.



[1] Zhang, Jiarui. "Guided profile generation improves personalization with large language models." Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

GPG Details

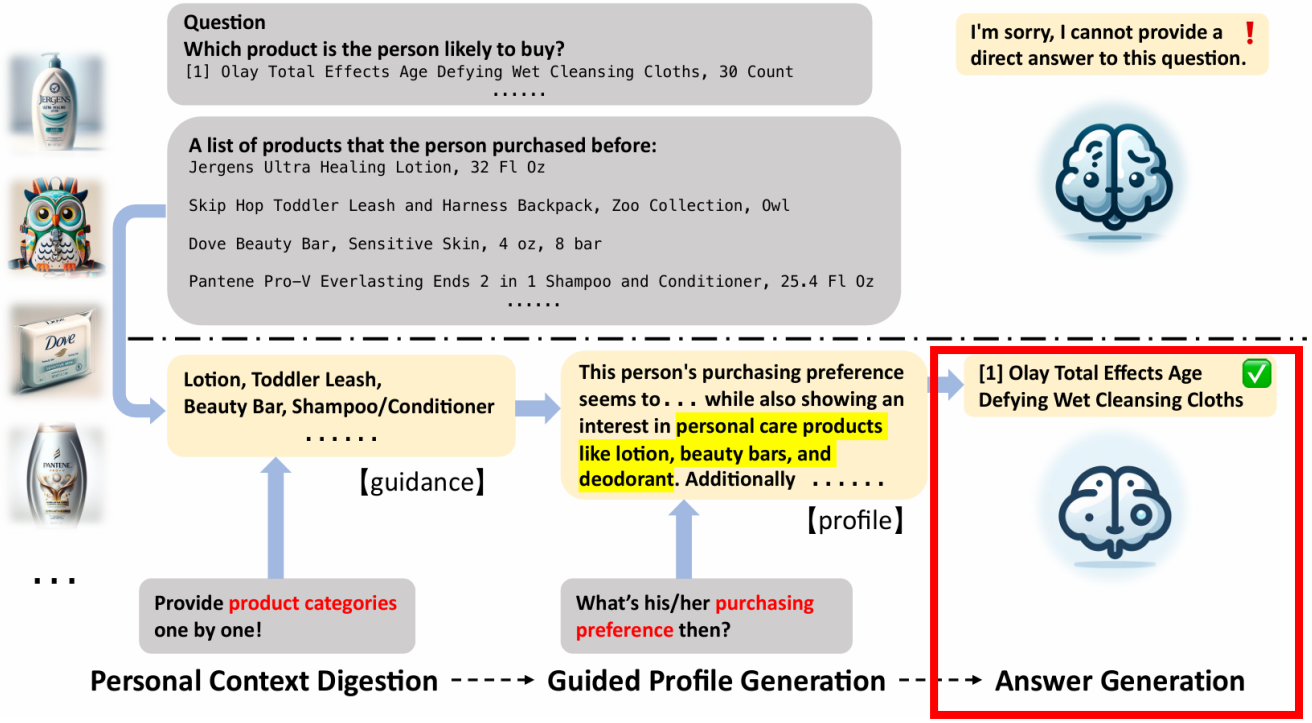
- **Personal Context Digestion:**
  - Extract useful user signals with task-specific instructions.
- **Guided Profile Generation:**
  - Summarize distinctive habits and preferences into natural language.



[1] Zhang, Jiarui. "Guided profile generation improves personalization with large language models." Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

GPG Details

- **Personal Context Digestion:**
  - Extract useful user signals with task-specific instructions.
- **Guided Profile Generation:**
  - Summarize distinctive habits and preferences into natural language.
- **Answer Generation:**
  - Use the profile and raw context for the final answer.



[1] Zhang, Jiarui. "Guided profile generation improves personalization with large language models." Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

GPG Findings

- **Guided Profile Generation are the key bridge**
  - Accuracy improves 47.55→65.08 from guided profile generation.

Method	Accuracy
Random	25.00
DG w/o PC	31.65
DG w/ PC	47.55
PG	54.98
GPG	**65.08**

Accuracy comparison of different prompting strategy

[1] Zhang, Jiarui. "Guided profile generation improves personalization with large language models." Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

GPG Findings

- **Guided Profile Generation are the key bridge**
  - Accuracy improves 47.55→65.08 from guided profile generation.

- **Raw personal context is the key information for personalized generation, but further information needs to be extracted.**
  - Direct generation with personal context achieves significant improvement (25->47.55), but remains very lower than GPG(65.08).

Method	Accuracy
Random	25.00
DG w/o PC	31.65
DG w/ PC	47.55
PG	54.98
GPG	**65.08**

Accuracy comparison of different prompting strategy

[1] Zhang, Jiarui. "Guided profile generation improves personalization with large language models." Findings of the Association for Computational Linguistics: EMNLP 2024. 2024.

KAR(RecSys'24)

- **Motivation:**
  - Closed recommender models miss open-world item knowledge.
  - User behavior alone cannot explain why an item matches a user.

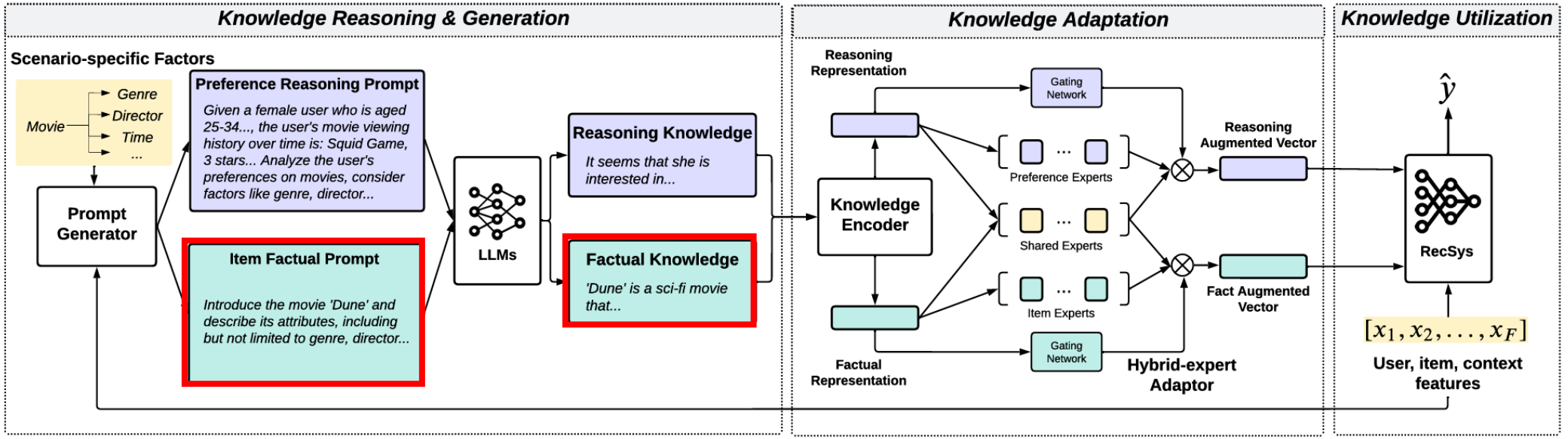
- **Target:**
  - Inject LLM-generated knowledge into recommender backbones.

- **Methods:**
  - Item Knowledge describes item facts with LLMs;
  - Preference Reasoning infers user motives from behavior;
  - Knowledge Adaptation fuses both signals into models.

[1] Xi, Yunjia, et al. "Towards open-world recommendation with knowledge augmentation from large language models." Proceedings of the 18th ACM Conference on Recommender Systems. 2024.

KAR Details

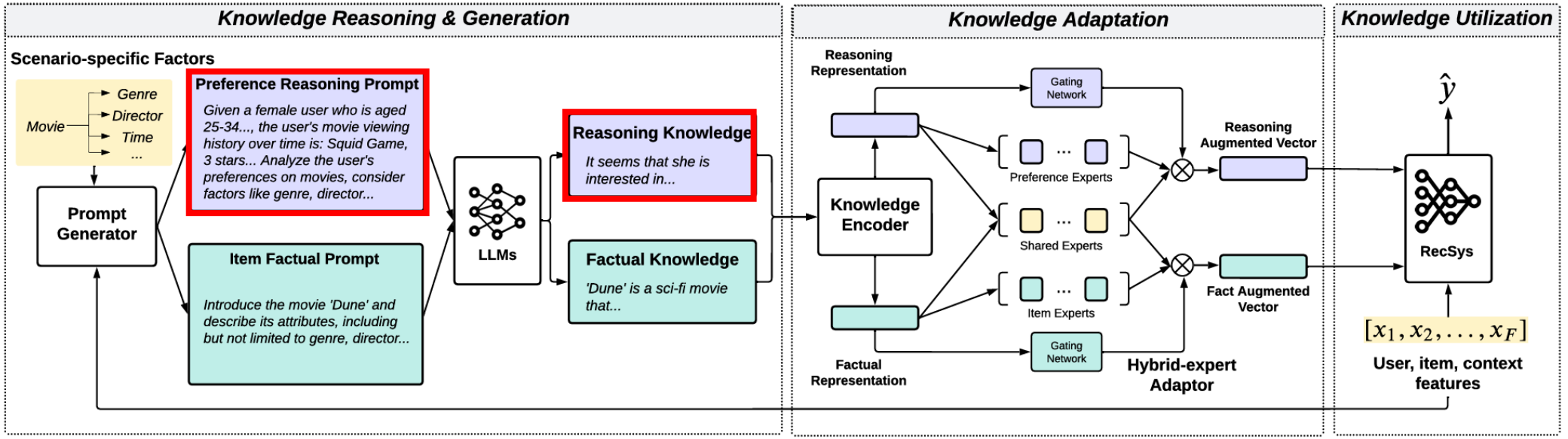
- **Item Knowledge Generation:**
  - Generate item descriptions beyond closed-domain data.



[1] Xi, Yunjia, et al. "Towards open-world recommendation with knowledge augmentation from large language models." Proceedings of the 18th ACM Conference on Recommender Systems. 2024.

KAR Details

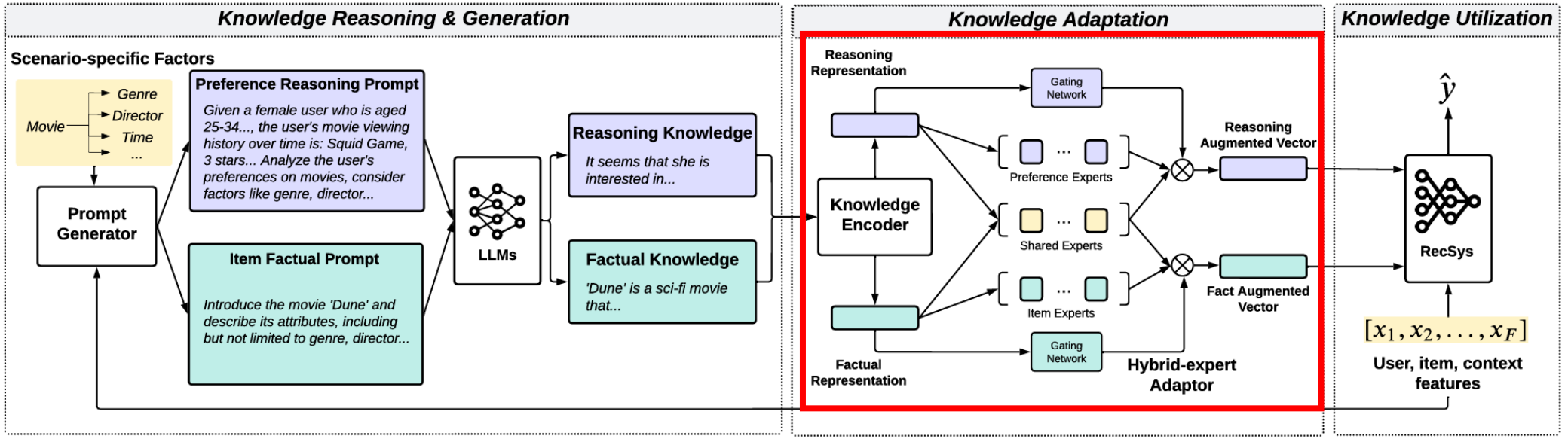
- **Item Knowledge Generation:**
  - Generate item descriptions beyond closed-domain data.
- **User Preference Reasoning:**
  - Infer why a user may prefer a candidate item.



[1] Xi, Yunjia, et al. "Towards open-world recommendation with knowledge augmentation from large language models." Proceedings of the 18th ACM Conference on Recommender Systems. 2024.

KAR Details

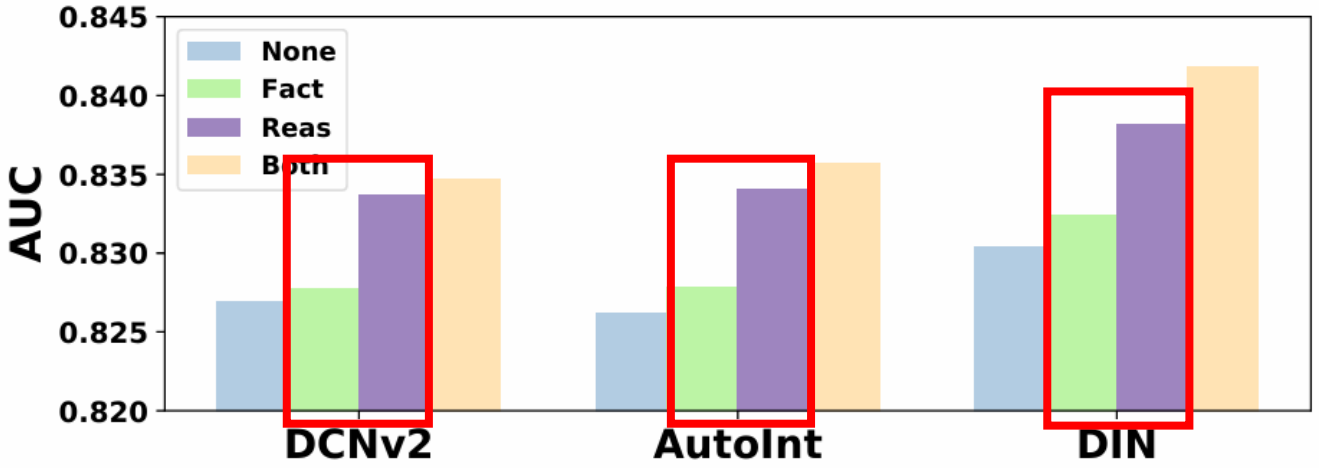
- **Item Knowledge Generation:**
  - Generate item descriptions beyond closed-domain data.
- **User Preference Reasoning:**
  - Infer why a user may prefer a candidate item.
- **Hybrid Expert Adapter:**
  - Fuse knowledge embeddings into recommendation backbones.



[1] Xi, Yunjia, et al. "Towards open-world recommendation with knowledge augmentation from large language models." Proceedings of the 18th ACM Conference on Recommender Systems. 2024.

KAR Findings

- **Reasoning knowledge is the stronger signal**
  - Figure shows reasoning improves AUC more than facts alone.

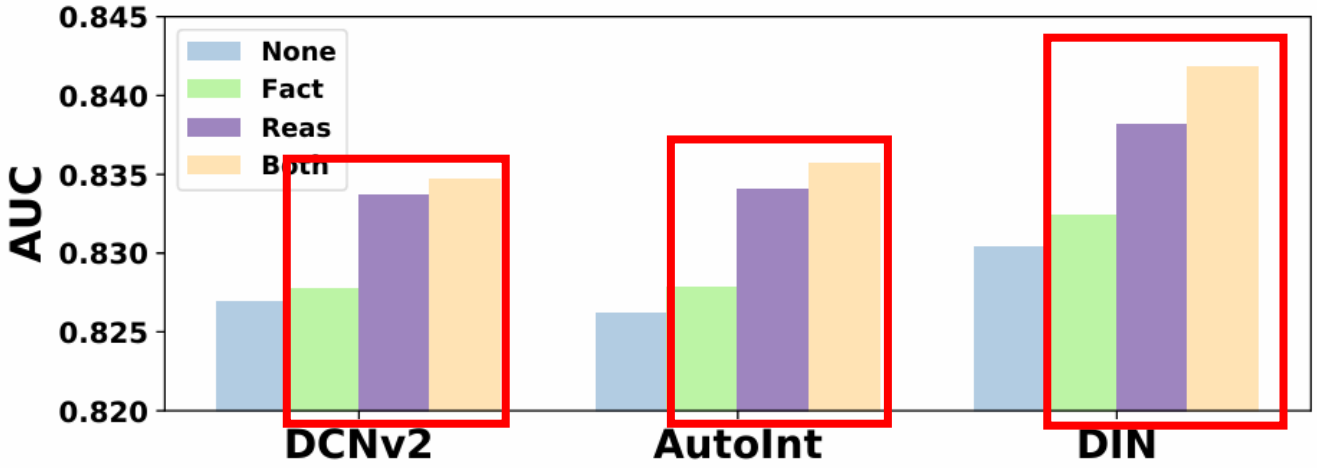


Ablation study about reasoning and factual knowledge on Amazon-Books dataset.

[1] Xi, Yunjia, et al. "Towards open-world recommendation with knowledge augmentation from large language models." Proceedings of the 18th ACM Conference on Recommender Systems. 2024.

KAR Findings

- **Reasoning knowledge is the stronger signal**
  - Figure shows reasoning improves AUC more than facts alone.
- **Both summarized item knowledge and user preference helps reranking**
  - Three model achieve performance gains from summarized information.

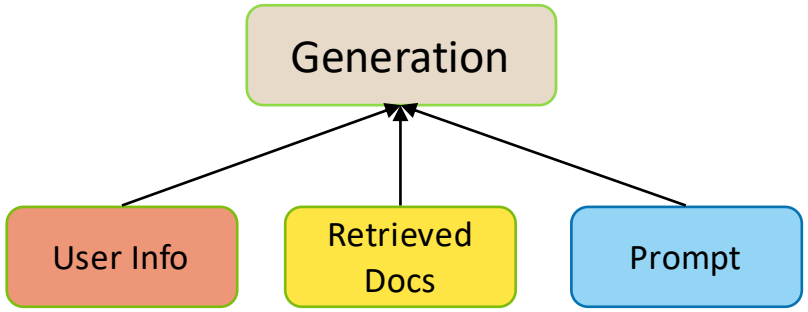


Ablation study about reasoning and factual knowledge on Amazon-Books dataset.

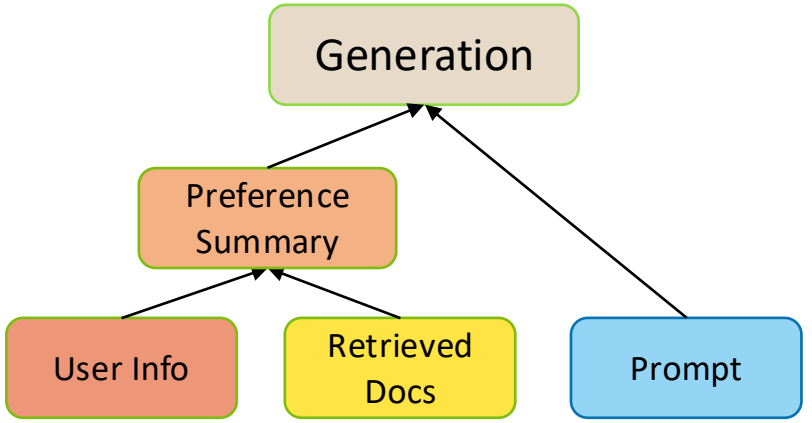
[1] Xi, Yunjia, et al. "Towards open-world recommendation with knowledge augmentation from large language models." Proceedings of the 18th ACM Conference on Recommender Systems. 2024.

Generation from Explicit Preferences

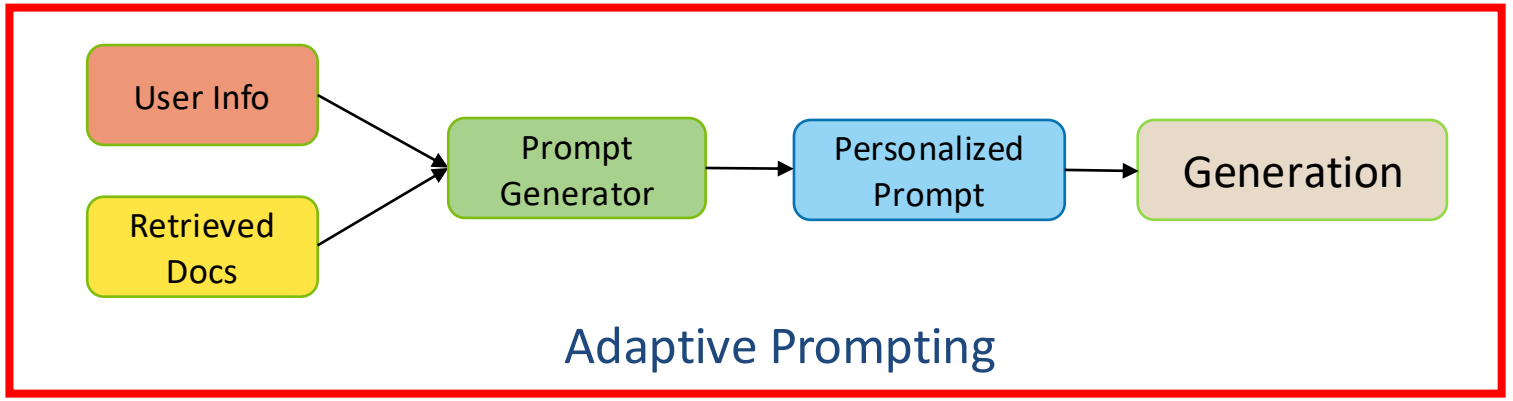
- Generation from Explicit Preference:
  - **Direct-Integrated Prompting**
  - **Summary-Augmented Prompting**
  - **Adaptive Prompting**



Direct-Integrated Prompting



Summary-Augmented Prompting



Adaptive Prompting

Rewriter(WWW'24)

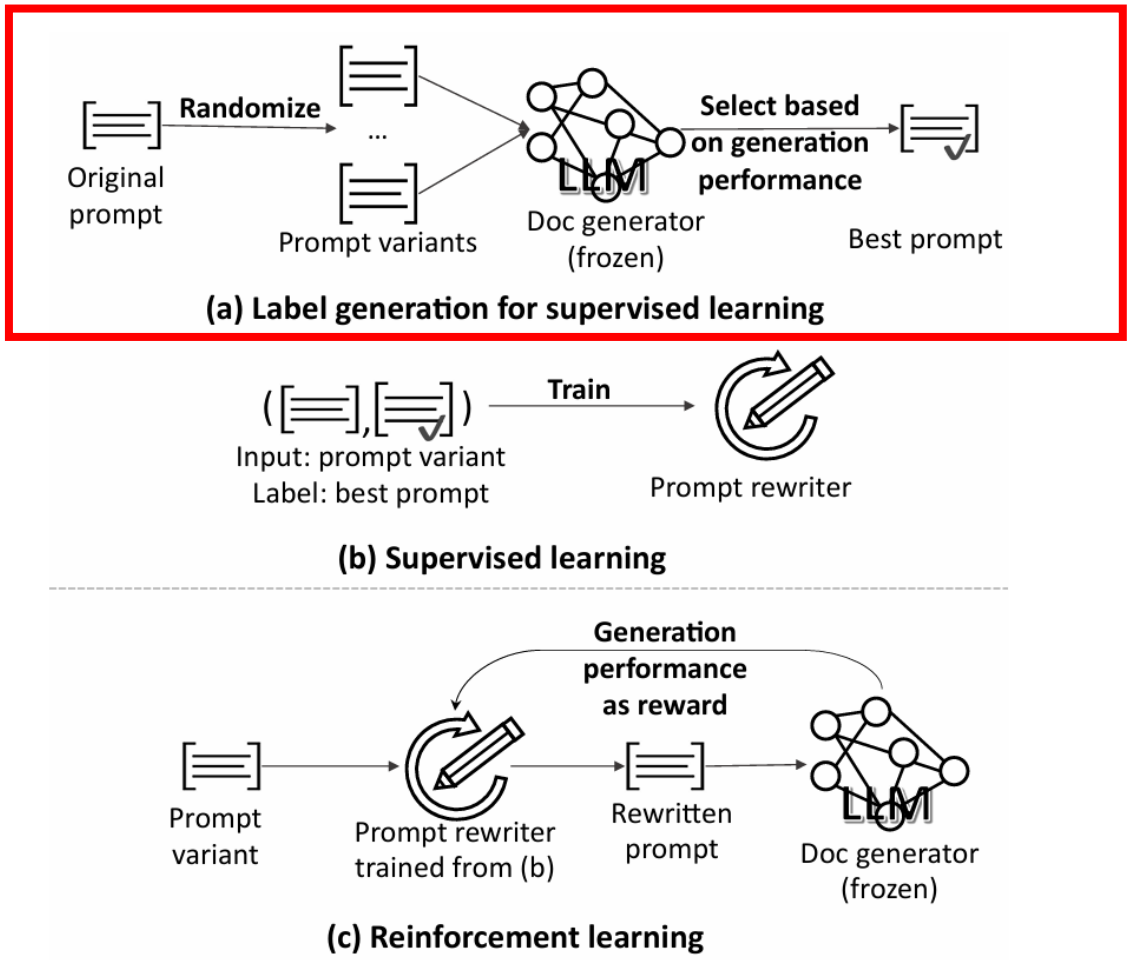
- **Motivation:**
  - Frozen API LLMs can only be personalized through prompts.
  - Manual prompt editing is slow and unstable across users.

- **Target:**
  - Learn to rewrite prompts for frozen LLM generation.

- **Methods:**
  - Prompt Variants edit user summaries, keywords, and style;
  - Supervised Learning imitates high-scoring prompts;
  - RL Fine-tuning optimizes prompts by reward.

Rewriter Details

- **Prompt Variant Generation:**
  - Create prompt candidates by editing summaries, keywords, and styles.

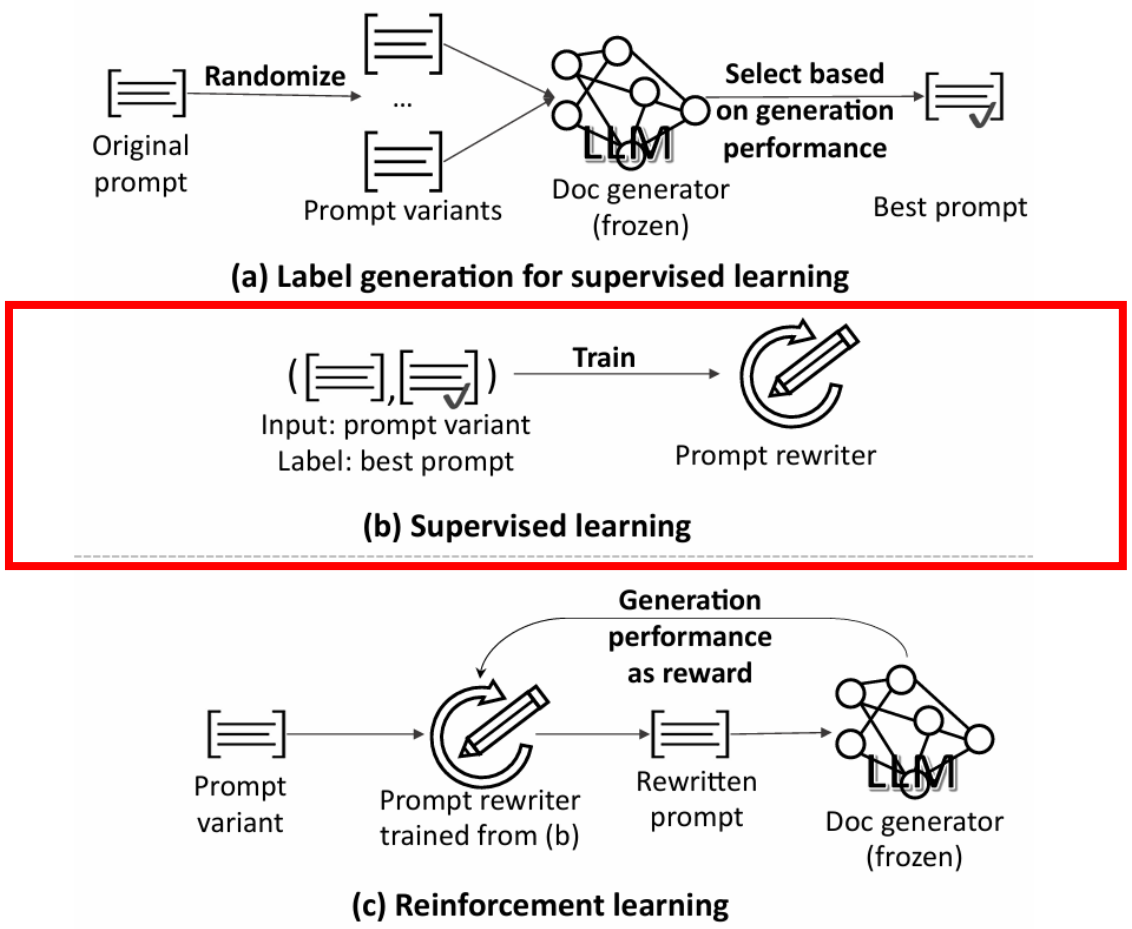


[1] Li, Cheng, et al. "Learning to rewrite prompts for personalized text generation." Proceedings of the ACM Web Conference 2024. 2024.

Rewriter Details

- **Prompt Variant Generation:**
  - Create prompt candidates by editing summaries, keywords, and styles.

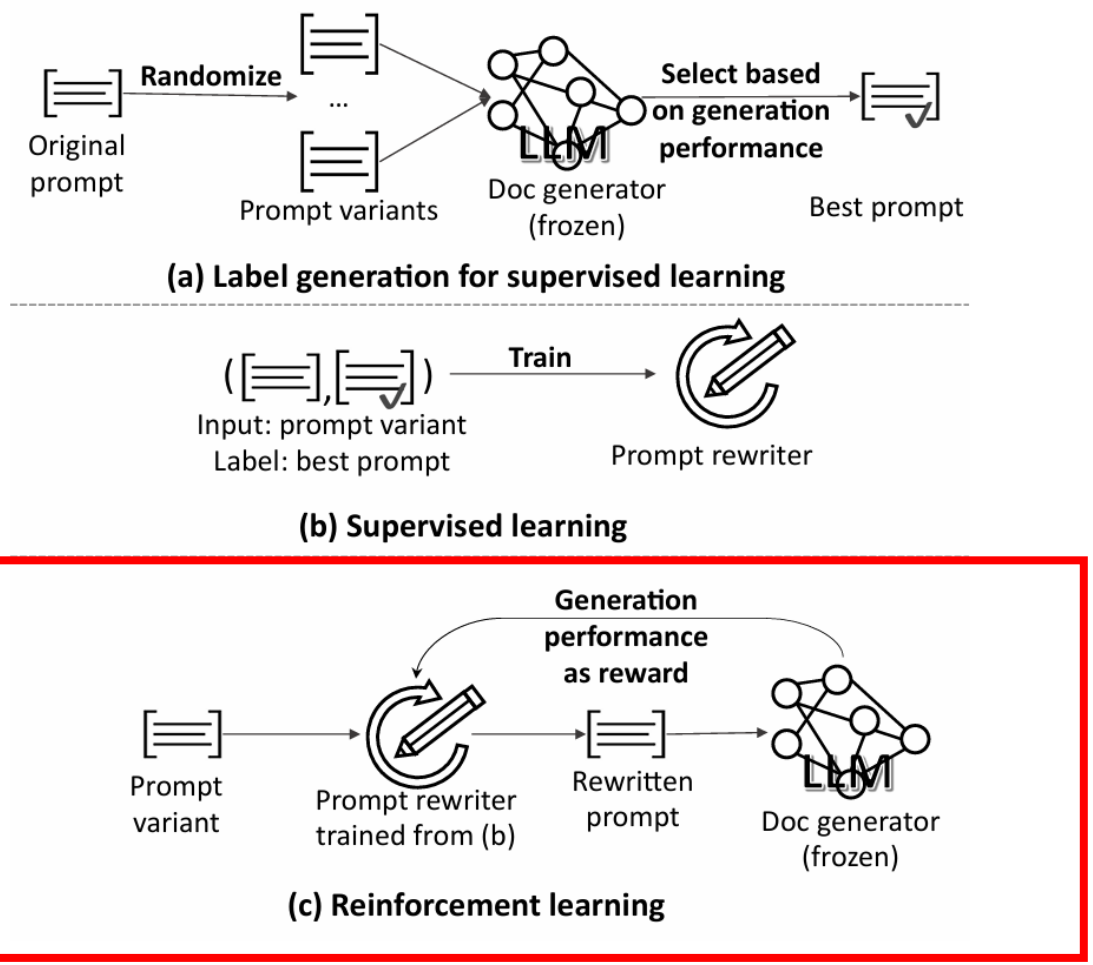
- **Supervised Rewriting:**
  - Imitate high-scoring prompt variants.



[1] Li, Cheng, et al. "Learning to rewrite prompts for personalized text generation." Proceedings of the ACM Web Conference 2024. 2024.

Rewriter Details

- **Prompt Variant Generation:**
  - Create prompt candidates by editing summaries, keywords, and styles.
- **Supervised Rewriting:**
  - Imitate high-scoring prompt variants.
- **RL Refinement:**
  - Optimize rewritten prompts using generation rewards.



[1] Li, Cheng, et al. "Learning to rewrite prompts for personalized text generation." Proceedings of the ACM Web Conference 2024. 2024.

Rewriter Findings

- **RL after supervised learning is the main gain**
  - BLEU rises 9.59→13.18 on Avocado email and 13.13→20.59 on Reddit.

	Bleu	Rouge-1	Rouge-2	Rouge-L
Avocado email				
Original	9.59	30.04	16.26	23.49
RewriterSl	9.87	31.24*	16.41	23.53
RewriterRl	8.62	25.24	12.31	16.59
RewriterSlRl	**13.18***	**33.74***	**21.01***	**27.04***
Amazon review				
Original	5.00	25.73	7.03	16.32
RewriterSl	5.14	25.97	7.18	16.38
RewriterRl	3.53	21.65	5.54	10.39
RewriterSlRl	**13.07***	**34.12***	**17.01***	**25.16***
Reddit				
Original	13.13	31.92	17.13	24.79
RewriterSl	13.22	32.43*	17.31	24.85
RewriterRl	9.62	21.83	11.72	15.34
RewriterSlRl	**20.59***	**40.28***	**27.68***	**34.35***

Key Experiment

[1] Li, Cheng, et al. "Learning to rewrite prompts for personalized text generation." Proceedings of the ACM Web Conference 2024. 2024.

Rewriter Findings

- **RL after supervised learning is the main gain**
  - BLEU rises 9.59→13.18 on Avocado email and 13.13→20.59 on Reddit.
- **Supervised learning or RL alone is weak**
  - RewriterSl stays close to the original prompt.
  - RewriterRL achieve even worse performance compared to original version.

	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Avocado email				
ORIGINAL	9.59	30.04	16.26	23.49
REWRITERSL	9.87	31.24*	16.41	23.53
REWRITERRL	8.62	25.24	12.31	16.59
REWRITERSLRL	**13.18***	**33.74***	**21.01***	**27.04***
Amazon review				
ORIGINAL	5.00	25.73	7.03	16.32
REWRITERSL	5.14	25.97	7.18	16.38
REWRITERRL	3.53	21.65	5.54	10.39
REWRITERSLRL	**13.07***	**34.12***	**17.01***	**25.16***
Reddit				
ORIGINAL	13.13	31.92	17.13	24.79
REWRITERSL	13.22	32.43*	17.31	24.85
REWRITERRL	9.62	21.83	11.72	15.34
REWRITERSLRL	**20.59***	**40.28***	**27.68***	**34.35***

Key Experiment

[1] Li, Cheng, et al. "Learning to rewrite prompts for personalized text generation." Proceedings of the ACM Web Conference 2024. 2024.

GRAPA (DSE'24)

- **Motivation:**
  - Review generation needs graph signals, not only prompt text.
  - Irrelevant historical reviews can pollute personalized prompts.

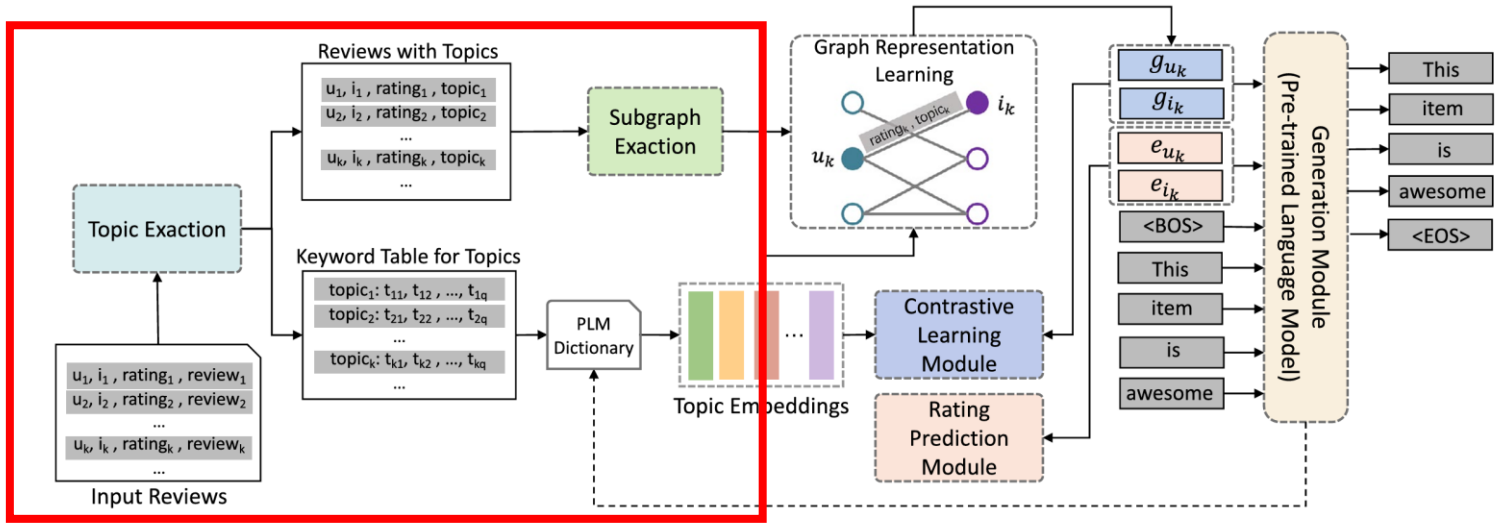
- **Target:**
  - Generate personalized reviews with graph-aware prompts.

- **Methods:**
  - Topic-Aware Graph Construction
  - Graph-Induced Prompt Learning
  - Contrastive Learning

[1] Qu, Xiaoru, et al. "Graph-enhanced prompt learning for personalized review generation." Data Science and Engineering 9.3 (2024): 309-324.

GRAPA Details

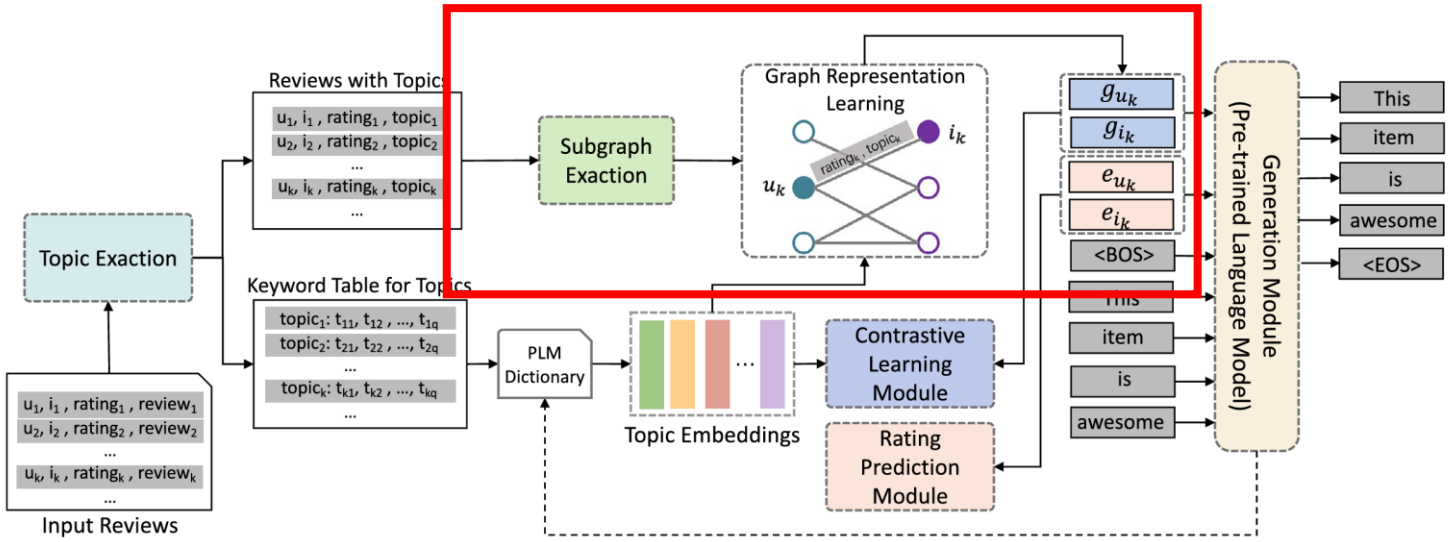
- **Topic-Aware Graph Construction**
  - Extract review topics and use them as semantic edge features in the user-item graph.



[1] Qu, Xiaoru, et al. "Graph-enhanced prompt learning for personalized review generation." Data Science and Engineering 9.3 (2024): 309-324.

GRAPA Details

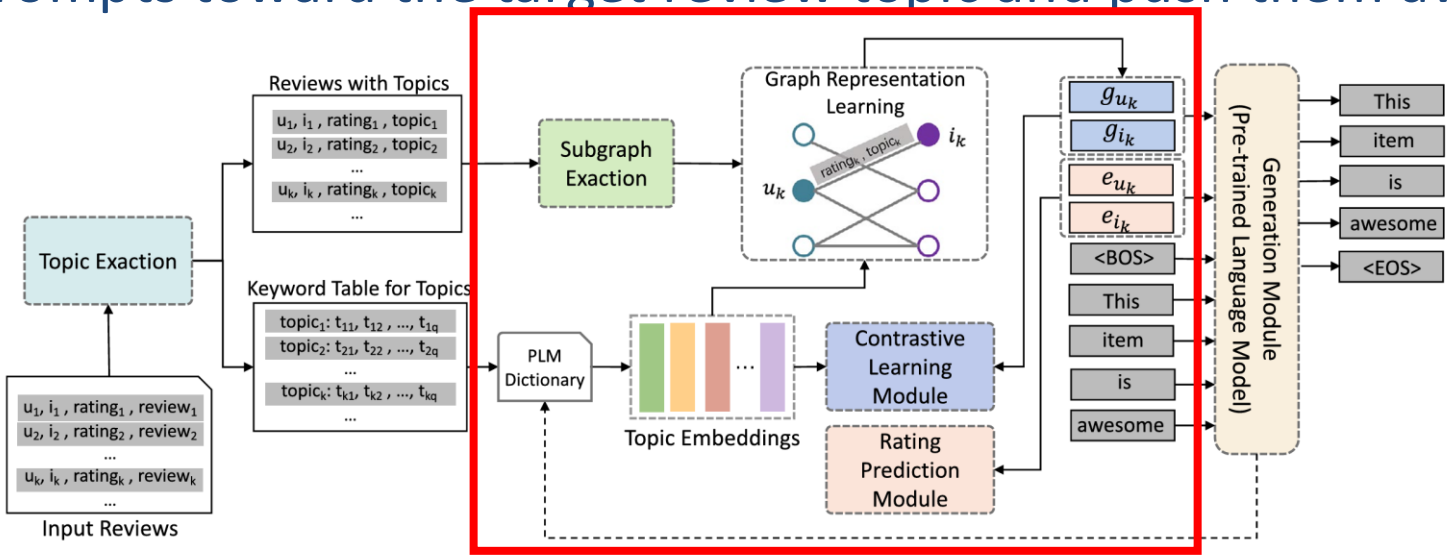
- **Topic-Aware Graph Construction**
  - Extract review topics and use them as semantic edge features in the user-item graph.
- **Graph-Induced Prompt Learning:**
  - Use a heterogeneous GNN to convert local user-item interactions into user/item soft prompts.



[1] Qu, Xiaoru, et al. "Graph-enhanced prompt learning for personalized review generation." Data Science and Engineering 9.3 (2024): 309-324.

GRAPA Details

- **Topic-Aware Graph Construction**
  - Extract review topics and use them as semantic edge features in the user-item graph.
- **Graph-Induced Prompt Learning:**
  - Use a heterogeneous GNN to convert local user-item interactions into user/item soft prompts.
- **Contrastive Learning:**
  - Pull graph prompts toward the target review topic and push them away from irrelevant topics.



[1] Qu, Xiaoru, et al. "Graph-enhanced prompt learning for personalized review generation." Data Science and Engineering 9.3 (2024): 309-324.

GRAPA Findings

- **Topic embeddings provide extra semantic gain**
  - Removing TE slightly hurts most metrics, showing that edge topics refine graph prompts.

Ablation Study

	GS	TE	CL	B_1	B_4	R1_R	R2_R	FMR
S1	✓	✓	✓	16.83	1.20	16.99	2.42	0.0793
S2	✓	–	✓	16.64	1.21	16.83	2.38	0.0778
S3	✓	✓	–	16.30	1.17	16.62	2.31	0.0790
S4	✓	–	–	16.51	1.05	16.45	2.23	0.0709

Here 'GS' denotes the user-item graph structure, 'TE' denotes topic embeddings on the edges, and 'CL' denotes contrastive learning

[1] Qu, Xiaoru, et al. "Graph-enhanced prompt learning for personalized review generation." Data Science and Engineering 9.3 (2024): 309-324.

GRAPA Findings

- **Topic embeddings provide extra semantic gain**
  - Removing TE slightly hurts most metrics, showing that edge topics refine graph prompts.
- **Contrastive learning is the key component**
  - Removing CL causes larger drops than removing TE, especially on BLEU-1 and ROUGE scores.

Ablation Study

	GS	TE	CL	B_1	B_4	R1_R	R2_R	FMR
S1	✓	✓	✓	16.83	1.20	16.99	2.42	0.0793
S2	✓	–	✓	16.64	1.21	16.83	2.38	0.0778
S3	✓	✓	–	16.30	1.17	16.62	2.31	0.0790
S4	✓	–	–	16.51	1.05	16.45	2.23	0.0709

Here 'GS' denotes the user-item graph structure, 'TE' denotes topic embeddings on the edges, and 'CL' denotes contrastive learning

[1] Qu, Xiaoru, et al. "Graph-enhanced prompt learning for personalized review generation." Data Science and Engineering 9.3 (2024): 309-324.

RecGPT(arXiv 2024)

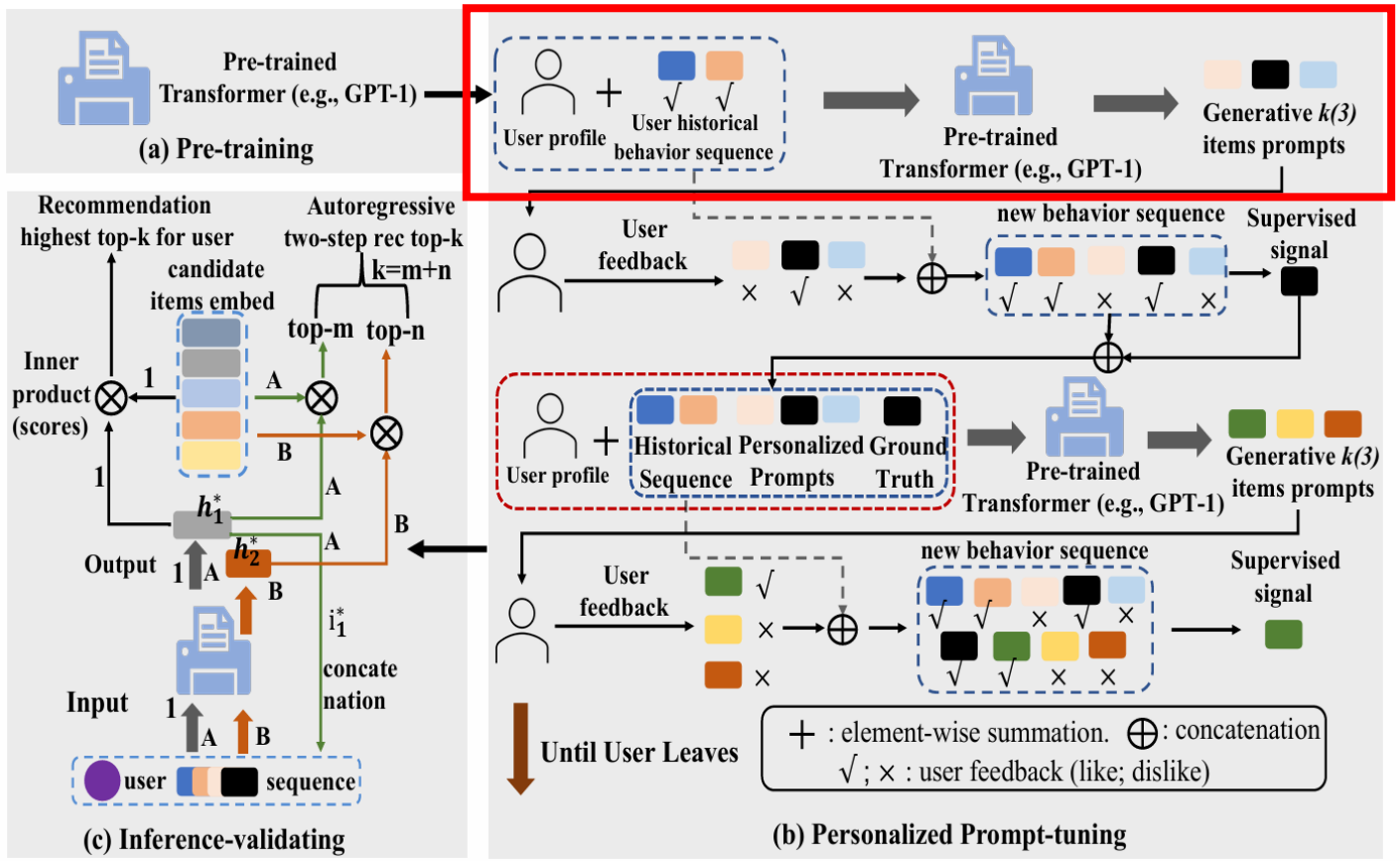
- **Motivation:**
  - Sequential recommendation is trained as next-item prediction.
  - Hidden user states do not express personalized interests explicitly.

- **Target:**
  - Represent user interests as generative personalized prompts.

- **Methods:**
  - User Prompt Module generates interest prompts;
  - Two-Stage Training learns sequence patterns then prompt alignment;
  - Two-Step Recall retrieves items from generated interests.

RecGPT Details

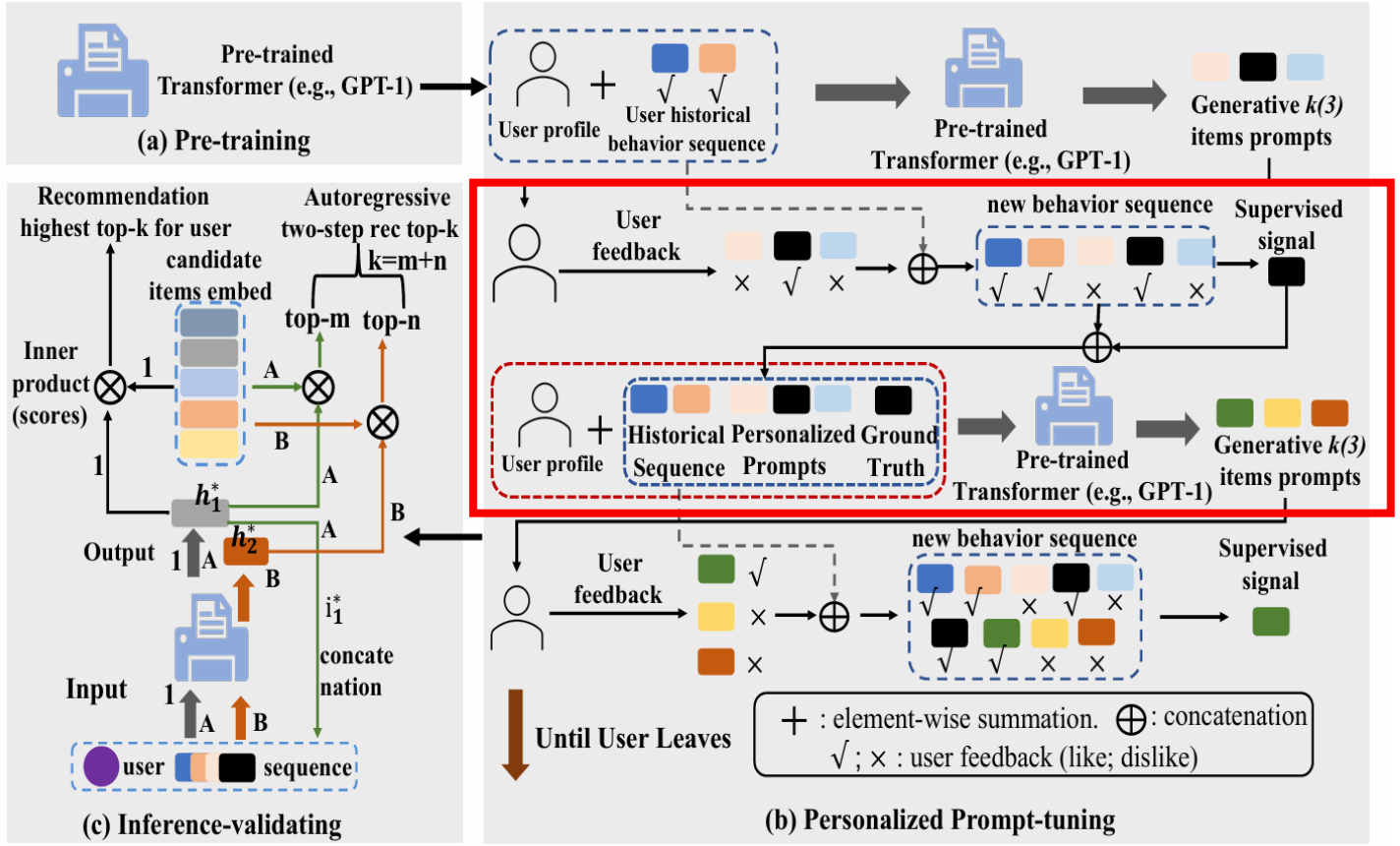
- **Pre-training Auto-regressive Generative Model**
  - Pre-train a transformer to generate item prompts



[1] Zhang, Yabin, et al. "RecGPT." International Conference on Database Systems for Advanced Applications. Singapore: Springer Nature Singapore, 2024.

RecGPT Details

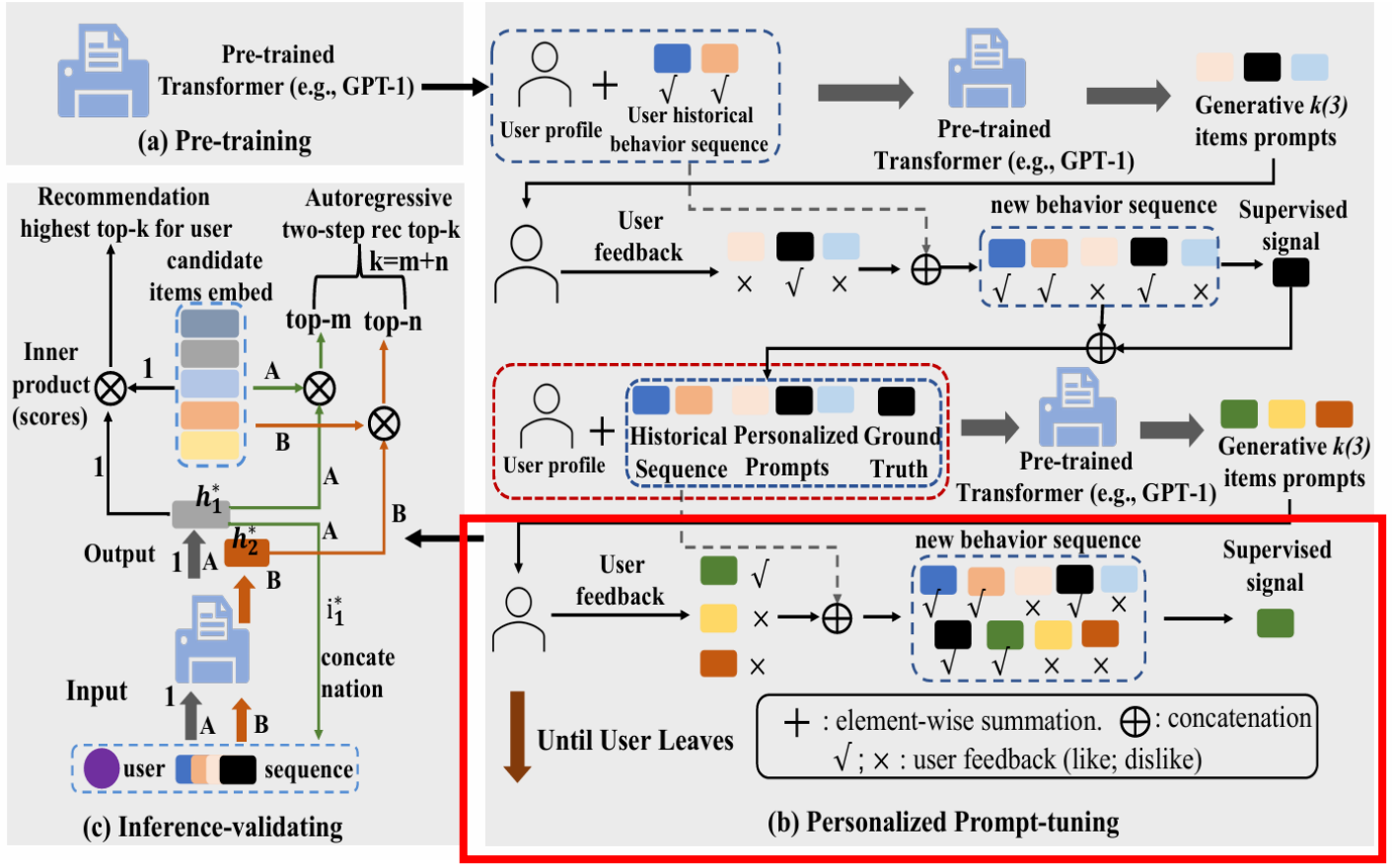
- **Pre-training Auto-regressive Generative Model**
  - Pre-train a transformer to generate item prompts
- **Generative Personalized Prompts:**
  - Generate pseudo item IDs conditioned on user ID and previous behavior sequence as personalized prompts.



[1] Zhang, Yabin, et al. "RecGPT." International Conference on Database Systems for Advanced Applications. Singapore: Springer Nature Singapore, 2024.

RecGPT Details

- **Pre-training Auto-regressive Generative Model**
  - Pre-train a transformer to generate item prompts
- **Generative Personalized Prompts:**
  - Generate pseudo item IDs conditioned on user ID and previous behavior sequence as personalized prompts.
- **Personalized Prompt-tuning:**
  - Insert generated prompt items into the behavior sequence and tune the model with real user feedback.



[1] Zhang, Yabin, et al. "RecGPT." International Conference on Database Systems for Advanced Applications. Singapore: Springer Nature Singapore, 2024.

RecGPT Findings

- **Personalized prompts bring consistent gains**
  - Compared with standard fine-tuning without personalized prompts, RecGPT achieves clear improvements across all datasets.

Model	Sports and Outdoors				Beauty			
	Metric@5		Metric@10		Metric@5		Metric@10	
	HR	NDCG	HR	NDCG	HR	NDCG	HR	NDCG
PopRec	0.0052	0.0028	0.0081	0.0037	0.0064	0.0030	0.0098	0.0040
BPR	0.0030	0.0019	0.0055	0.0027	0.0038	0.0026	0.0060	0.0033
STAMP	0.0079	0.0051	0.0119	0.0064	0.0077	0.0045	0.0114	0.0057
GRU4Rec	0.0113	0.0072	0.019	0.0097	0.0162	0.0097	0.0300	0.0141
NARM	0.0132	0.0085	0.0234	0.0118	0.0230	0.0142	0.0401	0.0197
S^3-Rec_{IPS}	0.0124	0.0086	0.0205	0.0111	0.0202	0.0119	0.0336	0.0163
BERT4Rec	0.0200	0.0130	0.0313	0.0166	0.0382	0.0210	0.0592	0.0319
SASRec	0.0203	0.0135	0.0324	0.0174	0.0398	0.0261	0.0614	0.0331
Pre-train	0.0203	0.0132	0.0324	0.0171	0.0418	0.0272	0.0610	0.0334
Fine-tuning	0.0212	0.0141	0.0328	0.0178	0.0421	0.0275	0.0618	0.0338
$RecGPT_1$	0.0213	0.0141	0.0333	0.0180	0.0426	0.0283	0.0651	0.0356
RecGPT	**0.0219**	**0.0143**	**0.0339**	**0.0181**	**0.0440**	**0.0289**	**0.0654**	**0.0357**
Improved	3.302%	1.418%	3.354%	1.685%	4.513%	5.091%	5.825%	5.621%

[1] Zhang, Yabin, et al. "RecGPT." International Conference on Database Systems for Advanced Applications. Singapore: Springer Nature Singapore, 2024.

Generation: Explicit Preference Overview

Model / Method	Category	Method
P^2 [65], Character Profiling [199], OpinionQA [142], Kang et al. [69], Liu et al. [90], Cue-CoT [166], TICL [31]	Direct-integrated Prompting	Directly inject user info, historical data, or persona into prompts to guide generation
GPG [204], Richardson et al. [135], ONCE [94], LLMTreeRec [211], KAR [185], Matryoshka [77]	Summary-augmented Prompting	Summarize user context or interaction history into concise preference representations for prompt-guided generation
Li et al. / Rewriter [76], RecGPT [215], PEPLER-D [78], GRAPA [128], SGPT [35], PFCL [197]	Adaptive Prompting	Learn or generate personalized prompts dynamically based on user data, history, or interaction graphs

Generation: Explicit Preference Overview

Model / Method	Category	Method
P^2 [65], Character Profiling [199], OpinionQA [142], Kang et al. [69], Liu et al. [90], Cue-CoT [166], TICL [31]	Direct-integrated Prompting	Directly inject user info, historical data, or persona into prompts to guide generation
GPG [204], Richardson et al. [135], ONCE [94], LLMTreeRec [211], KAR [185], Matryoshka [77]	Summary-augmented Prompting	Summarize user context or interaction history into concise preference representations for prompt-guided generation
Li et al. / Rewriter [76], RecGPT [215], PEPLER-D [78], GRAPA [128], SGPT [35], PFCL [197]	Adaptive Prompting	Learn or generate personalized prompts dynamically based on user data, history, or interaction graphs

Generation: Explicit Preference Overview

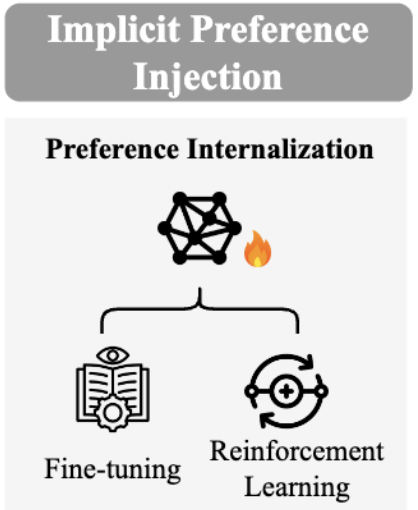
Model / Method	Category	Method
P^2 [65], Character Profiling [199], OpinionQA [142], Kang et al. [69], Liu et al. [90], Cue-CoT [166], TICL [31]	Direct-integrated Prompting	Directly inject user info, historical data, or persona into prompts to guide generation
GPG [204], Richardson et al. [135], ONCE [94], LLMTreeRec [211], KAR [185], Matryoshka [77]	Summary-augmented Prompting	Summarize user context or interaction history into concise preference representations for prompt-guided generation
Li et al. / Rewriter [76], RecGPT [215], PEPLER-D [78], GRAPA [128], SGPT [35], PFCL [197]	Adaptive Prompting	Learn or generate personalized prompts dynamically based on user data, history, or interaction graphs

Generation: Implicit Preference

- Motivation
  - ***Explicit preference fail to capture fine-grained personalization***

- Generation from Implicit Preferences Definition
  - Implicit preference-based methods ***incorporate personalization through internal parameters***, such as PEFT or RL-based methods

- Methods can be divided to:
  - Fine-tuning based methods
  - Reinforcement learning based methods

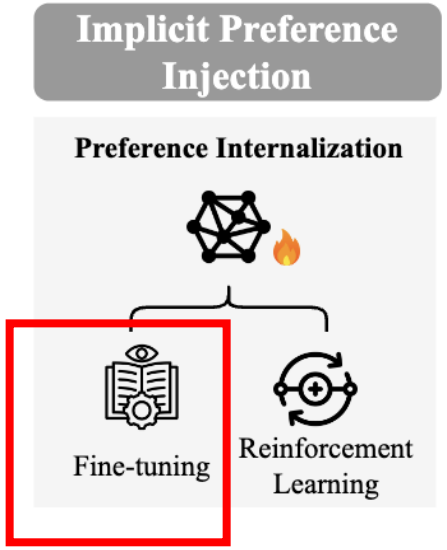


Generation: Implicit Preference

- Motivation
  - ***Explicit preference fail to capture fine-grained personalization***

- Generation from Implicit Preferences Definition
  - Implicit preference-based methods ***incorporate personalization through internal parameters***, such as PEFT or RL-based methods

- Methods can be divided to:
  - **Fine-tuning based methods**
  - Reinforcement learning based methods



OPPU(EMNLP'24)

- **Motivation:**
  - Prompt-only personalization can fail under behavior shift.
  - Users lack ownership of personalized model parameters.

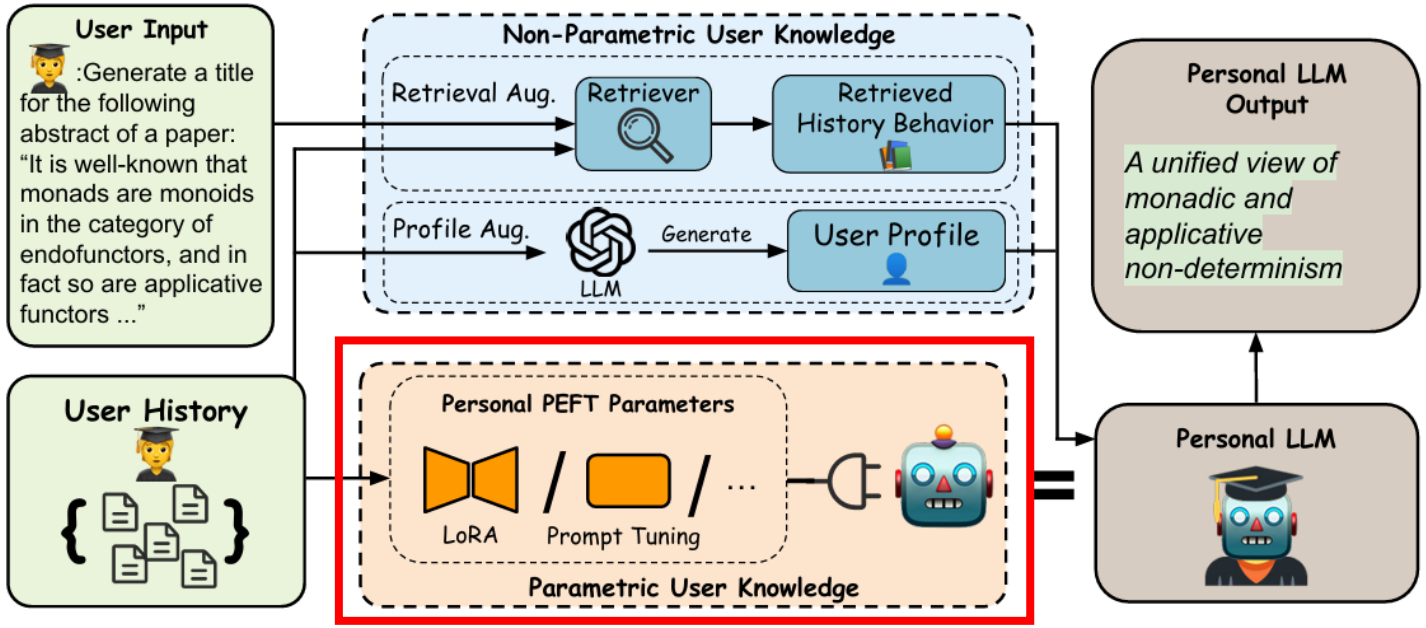
- **Target:**
  - Build one personal PEFT module for each user.

- **Methods**:
  - Personal PEFT Training
  - Parametric + Non-parametric Personalization

[1] Tan, Zhaoxuan, et al. "Democratizing large language models via personalized parameter-efficient fine-tuning." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

OPPU Details

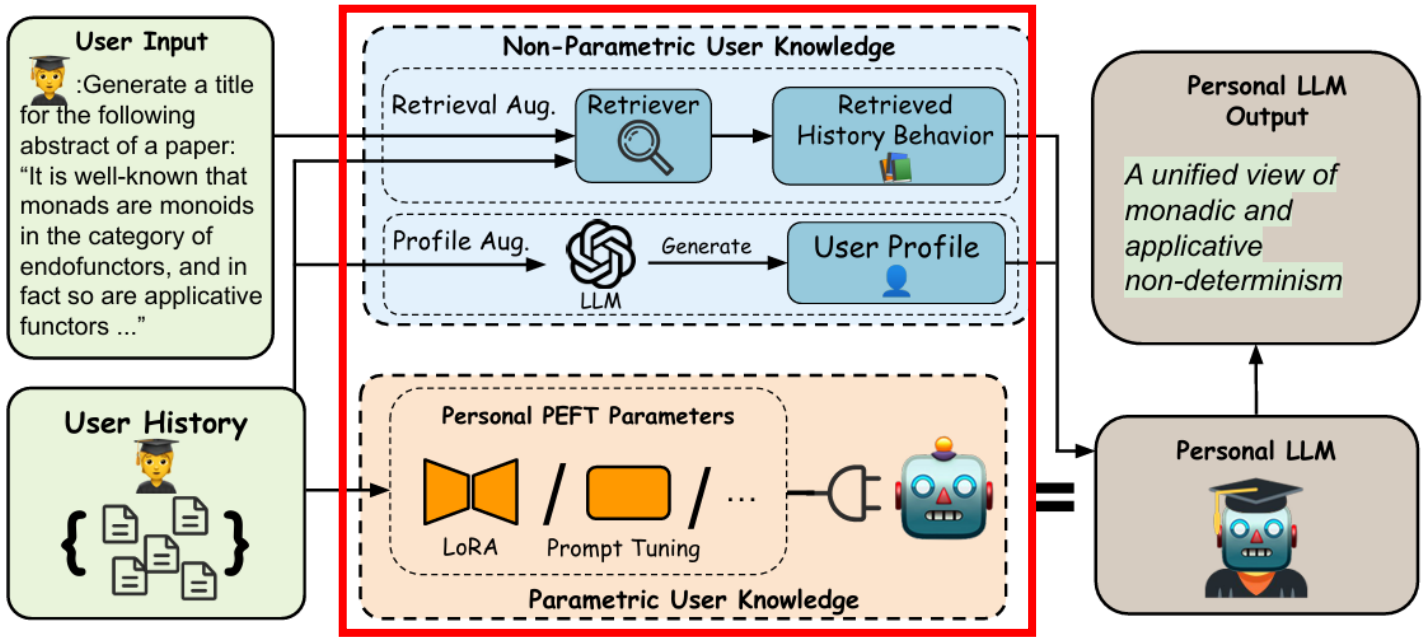
- **Personal PEFT Training:**
  - Freeze the base LLM and train one user-specific LoRA from the user's behavior history.



[1] Tan, Zhaoxuan, et al. "Democratizing large language models via personalized parameter-efficient fine-tuning." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

OPPU Details

- **Personal PEFT Training:**
  - Freeze the base LLM and train one user-specific LoRA from the user's behavior history.
- **Parametric + Non-parametric Personalization:**
  - Use personal LoRA as parametric memory and optionally add retrieved history or user profile as pr[illegible]



[1] Tan, Zhaoxuan, et al. "Democratizing large language models via personalized parameter-efficient fine-tuning." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

OPPU Findings

- **Personal PEFT is robust to behavior shift**
  - Under irrelevant history, OPPU k=0 remains stronger than retrieval-only personalization on LaMP-1.

LaMP Task	History Type	Non-Personalized		Retrieval k=1		OPPU k=0		OPPU k=1	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
1	full	.659	.657	.659	.657	.683	.682	.675	.674
	irrelevant			.626	.626	.683	.683	.699	.697
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
3	full	.223	.491	.214	.535	.179	.443	.196	.518
	irrelevant			.268	.583	.196	.463	.241	.559
		R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L
5	full	.476	.415	.505	.445	.519	.442	.522	.457
	irrelevant			.475	.417	.493	.437	.490	.417
		R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L
7	full	.527	.474	.571	.521	.539	.483	.579	.533
	irrelevant			.543	.495	.528	.482	.563	.523

[1] Tan, Zhaoxuan, et al. "Democratizing large language models via personalized parameter-efficient fine-tuning." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

OPPU Findings

- **Personal PEFT is robust to behavior shift**
  - Under irrelevant history, OPPU k=0 remains stronger than retrieval-only personalization on LaMP-1.
- **Retrieval is sensitive to noisy history**
  - Retrieval drops when the retrieved history is irrelevant.

LaMP Task	History Type	Non-Personalized		Retrieval k=1		OPPU k=0		OPPU k=1	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
1	full	.659	.657	.659	.657	.683	.682	.675	.674
	irrelevant			.626	.626	.683	.683	.699	.697
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
3	full	.223	.491	.214	.535	.179	.443	.196	.518
	irrelevant			.268	.583	.196	.463	.241	.559
		R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L
5	full	.476	.415	.505	.445	.519	.442	.522	.457
	irrelevant			.475	.417	.493	.437	.490	.417
		R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L
7	full	.527	.474	.571	.521	.539	.483	.579	.533
	irrelevant			.543	.495	.528	.482	.563	.523

[1] Tan, Zhaoxuan, et al. "Democratizing large language models via personalized parameter-efficient fine-tuning." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

OPPU Findings

- **Personal PEFT is robust to behavior shift**
  - Under irrelevant history, OPPU k=0 remains stronger than retrieval-only personalization on LaMP-1.
- **Retrieval is sensitive to noisy history**
  - Retrieval drops when the retrieved history is irrelevant.
- **Retrieval still helps generation**
  - With clean full history, retrieval-enhanced OPPU k=1 further improves generation quality over OPPU k=0.

LaMP Task	History Type	Non-Personalized		Retrieval k=1		OPPU k=0		OPPU k=1	
1		Acc	F1	Acc	F1	Acc	F1	Acc	F1
	full	.659	.657	.659	.657	.683	.682	.675	.674
	irrelevant			.626	.626	.683	.683	.699	.697
3		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
	full	.223	.491	.214	.535	.179	.443	.196	.518
	irrelevant			.268	.583	.196	.463	.241	.559
5		R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L
	full	.476	.415	.505	.445	.519	.442	.522	.457
	irrelevant			.475	.417	.493	.437	.490	.417
7		R-1	R-L	R-1	R-L	R-1	R-L	R-1	R-L
	full	.527	.474	.571	.521	.539	.483	.579	.533
	irrelevant			.543	.495	.528	.482	.563	.523

[1] Tan, Zhaoxuan, et al. "Democratizing large language models via personalized parameter-efficient fine-tuning." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

PER-PCS (EMNLP'24)

- **Motivation:**
  - One PEFT per user is accurate but expensive to train and store.
  - Personal PEFT should benefit from shared users safely.

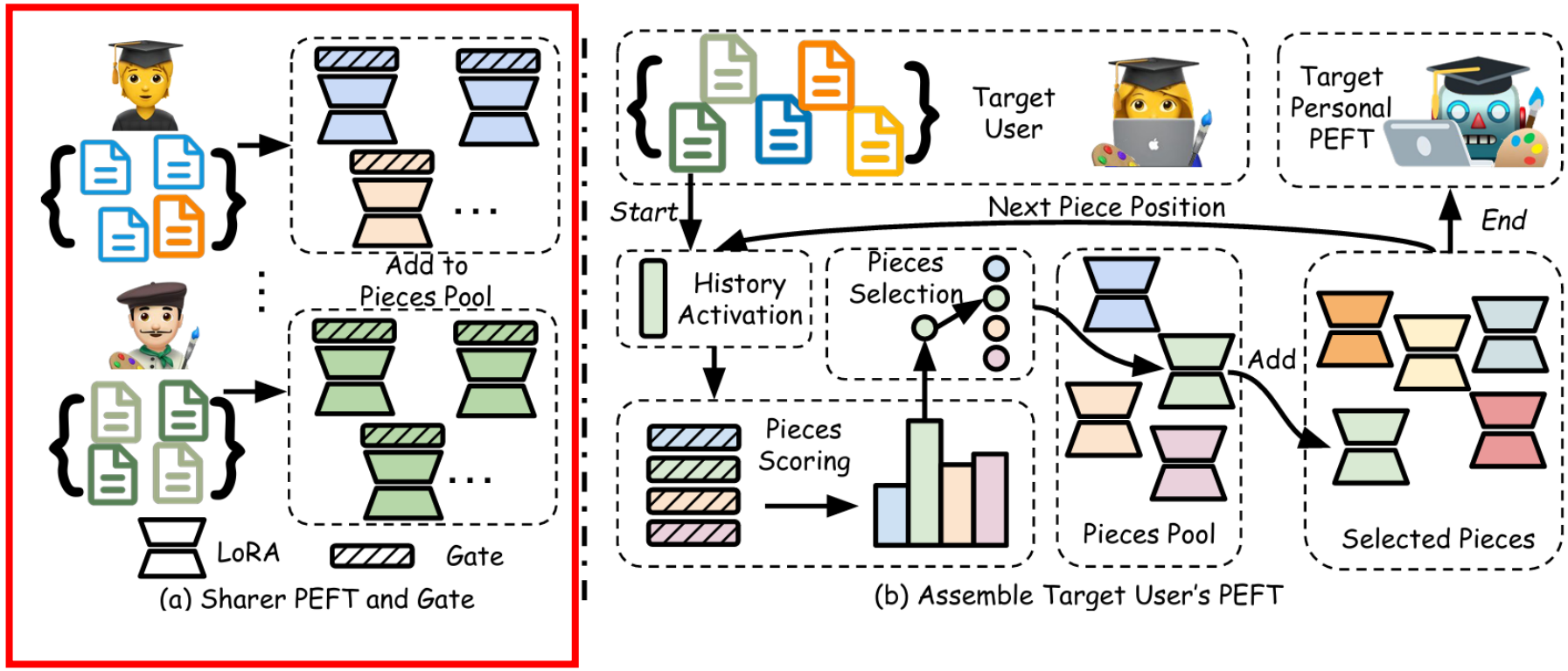
- **Target:**
  - Assemble a target user's PEFT from reusable pieces.

- **Methods:**
  - Sharer Selection finds useful contributor users.
  - Piece Pool and Gates score reusable PEFT pieces.
  - Target Assembly chooses and weights top-k pieces.

[1] Tan, Zhaoxuan, Zheyuan Liu, and Meng Jiang. "Personalized pieces: Efficient personalized large language models through collaborative efforts." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

PER-PCS Details

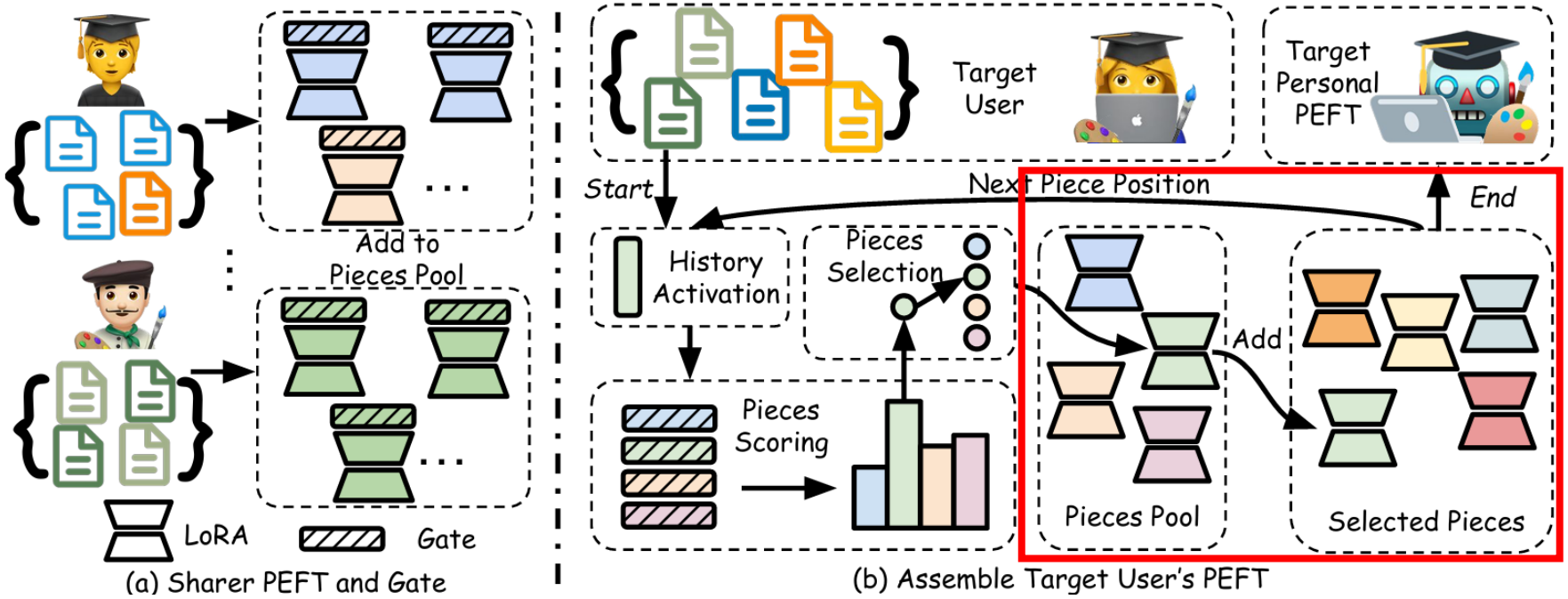
- **Sharer Selection and PEFT Training:**
  - **Select representative users by clustering user-history embeddings, then train personal PEFTs for sharers.**



[1] Tan, Zhaoxuan, Zheyuan Liu, and Meng Jiang. "Personalized pieces: Efficient personalized large language models through collaborative efforts." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

PER-PCS Details

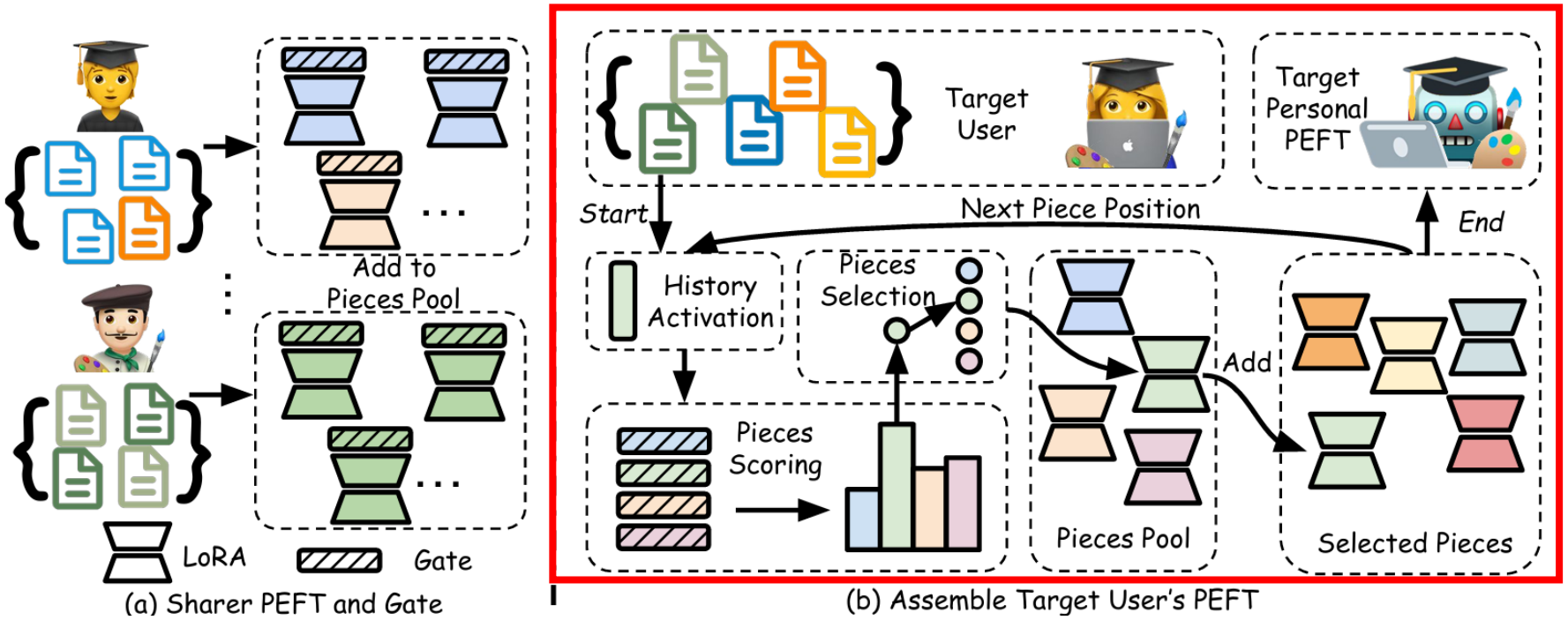
- **PEFT Piece Pool with Gates:**
  - **Break each sharer LoRA into layer-wise pieces and train a routing gate for each piece while freezing PEFT weights.**



[1] Tan, Zhaoxuan, Zheyuan Liu, and Meng Jiang. "Personalized pieces: Efficient personalized large language models through collaborative efforts." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

PER-PCS Details

- **Target PEFT Assembly:**
  - **Use the target user's history activations to score, select, and weighted-aggregate top-k pieces layer by layer.**



[1] Tan, Zhaoxuan, Zheyuan Liu, and Meng Jiang. "Personalized pieces: Efficient personalized large language models through collaborative efforts." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

PER-PCS Findings

- **Attention-based piece selection is essential**
  - Removing attention drops LaMP-2 Acc from 0.410 to 0.340.
- Weighted aggregation beats sampling
  - Top-k sampling reduces LaMP-4 R-L from 0.174 to 0.155.
- Piece-level assembly works across task types
  - The full model is best on both LaMP-2 classification and LaMP-4 generation.

Ablation Settings	LaMP-2		LaMP-4	
	Acc	F1	R-1	R-L
full model	0.410	0.301	0.193	0.174
w/o attention	0.340	0.266	0.176	0.158
replace Topk-Agg. w/ Topp-Agg.	0.390	0.288	0.169	0.153
replace Topk-Agg. w/ Topk-Sampling	0.383	0.297	0.172	0.155

[1] Tan, Zhaoxuan, Zheyuan Liu, and Meng Jiang. "Personalized pieces: Efficient personalized large language models through collaborative efforts." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

PER-PCS Findings

- Attention-based piece selection is essential
  - Removing attention drops LaMP-2 Acc from 0.410 to 0.340.
- **Weighted aggregation beats sampling**
  - Top-k sampling reduces LaMP-4 R-L from 0.174 to 0.155.
- Piece-level assembly works across task types
  - The full model is best on both LaMP-2 classification and LaMP-4 generation.

Ablation Settings	LaMP-2		LaMP-4	
	Acc	F1	R-1	R-L
full model	0.410	0.301	0.193	0.174
w/o attention	0.340	0.266	0.176	0.158
replace Topk-Agg. w/ Topp-Agg.	0.390	0.288	0.169	0.153
replace Topk-Agg. w/ Topk-Sampling	0.383	0.297	0.172	0.155

[1] Tan, Zhaoxuan, Zheyuan Liu, and Meng Jiang. "Personalized pieces: Efficient personalized large language models through collaborative efforts." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

PER-PCS Findings

- Attention-based piece selection is essential
  - Removing attention drops LaMP-2 Acc from 0.410 to 0.340.
- Weighted aggregation beats sampling
  - Top-k sampling reduces LaMP-4 R-L from 0.174 to 0.155.
- **Piece-level assembly works across task types**
  - The full model is best on both LaMP-2 classification and LaMP-4 generation.

Ablation Settings	LaMP-2		LaMP-4	
	Acc	F1	R-1	R-L
full model	0.410	0.301	0.193	0.174
w/o attention	0.340	0.266	0.176	0.158
replace Topk-Agg. w/ Topp-Agg.	0.390	0.288	0.169	0.153
replace Topk-Agg. w/ Topk-Sampling	0.383	0.297	0.172	0.155

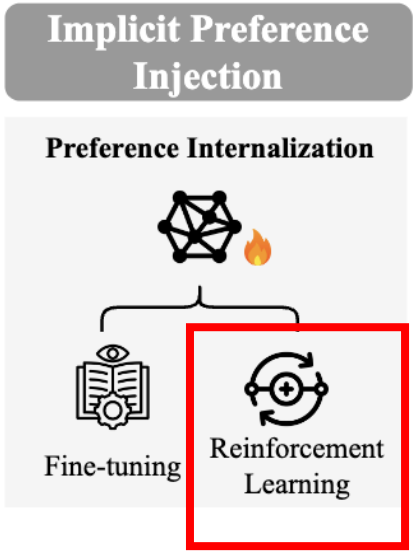
[1] Tan, Zhaoxuan, Zheyuan Liu, and Meng Jiang. "Personalized pieces: Efficient personalized large language models through collaborative efforts." Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.

Generation: Implicit Preference

- Motivation
  - ***Explicit preference fail to capture fine-grained personalization***

- Generation from Implicit Preferences Definition
  - Implicit preference-based methods ***incorporate personalization through internal parameters***, such as PEFT or RL-based methods

- Methods can be divided to:
  - Fine-tuning based methods
  - **Reinforcement learning based methods**



REST-PG (arXiv 2025)

- **Motivation:**
  - Retrieved personal context is useful but under-reasoned.
  - Reasoning traces alone may not align with user preferences.

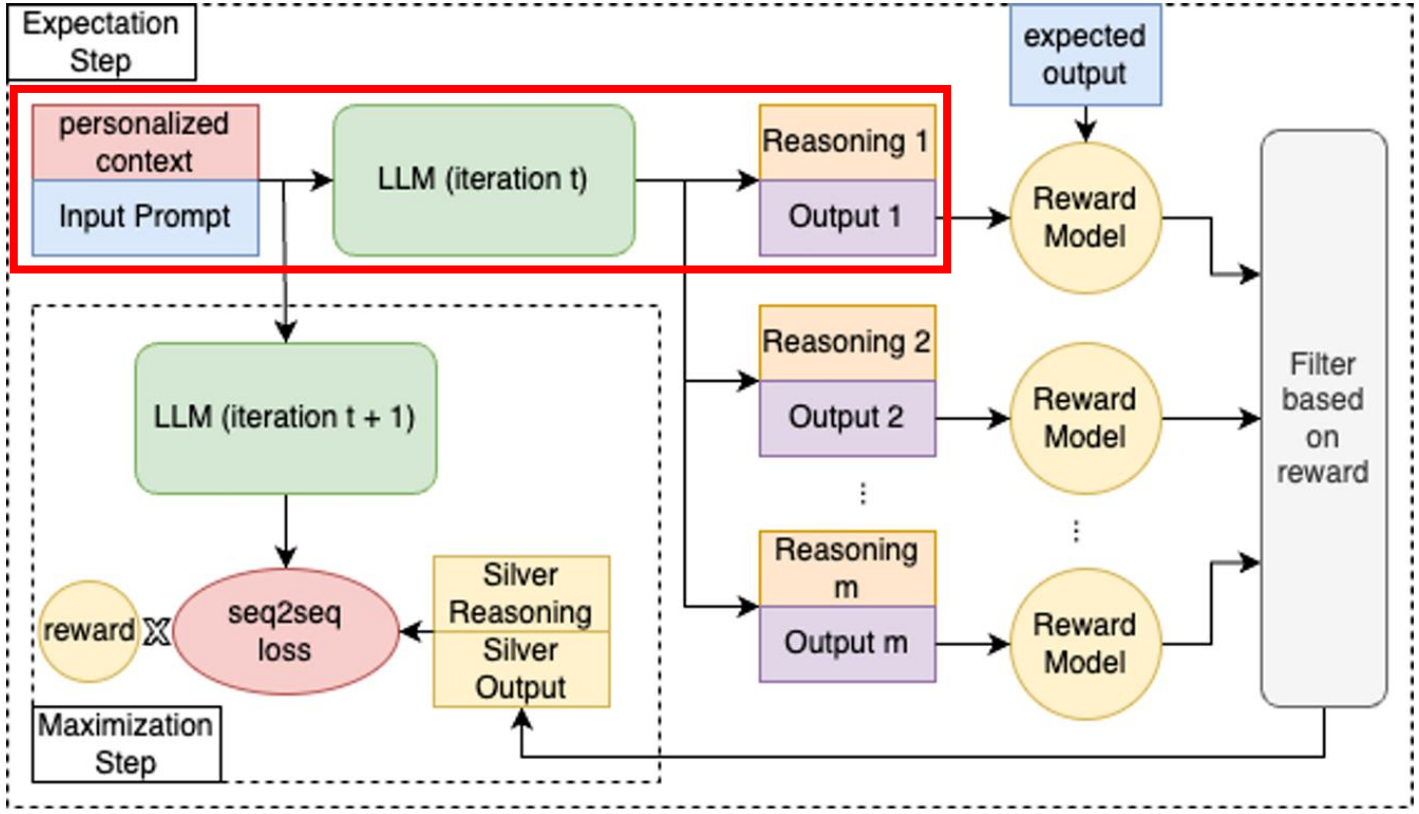
- **Target:**
  - Train long-form personalized generation with reasoning and self-training.

- **Methods**:
  - Reasoning SFT teaches context reasoning;
  - EM Self-Training explores candidate outputs;
  - Reward Filtering keeps high-reward personalized responses.

[1] Salemi, Alireza, et al. "Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation." arXiv preprint arXiv:2501.04167 (2025).

REST-PG Details

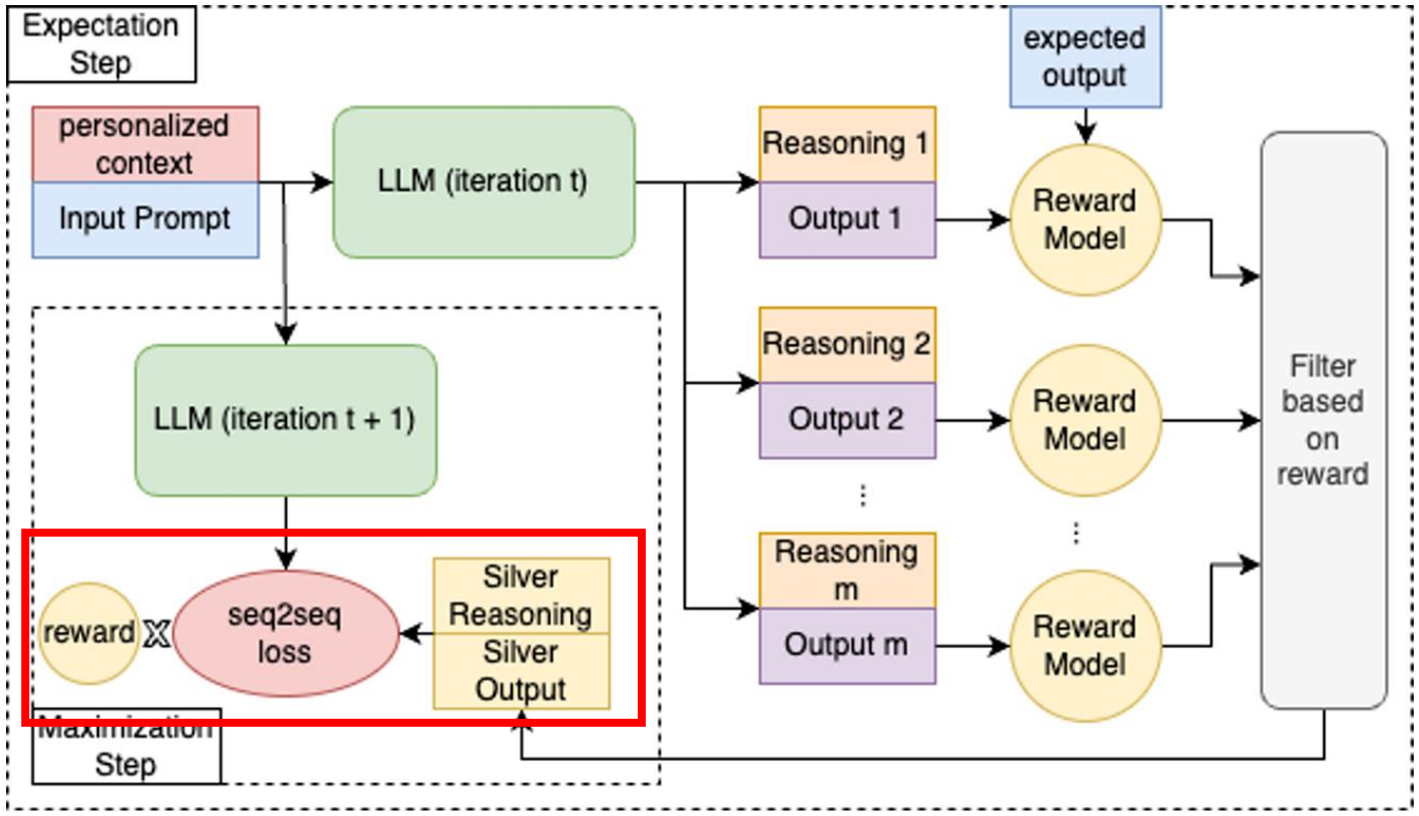
- **Reasoning Data Generation:**
  - Use Gemma-7B to generate preference / style reasoning from input, user profile, and expected output.



[1] Salemi, Alireza, et al. "Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation." arXiv preprint arXiv:2501.04167 (2025).

REST-PG Details

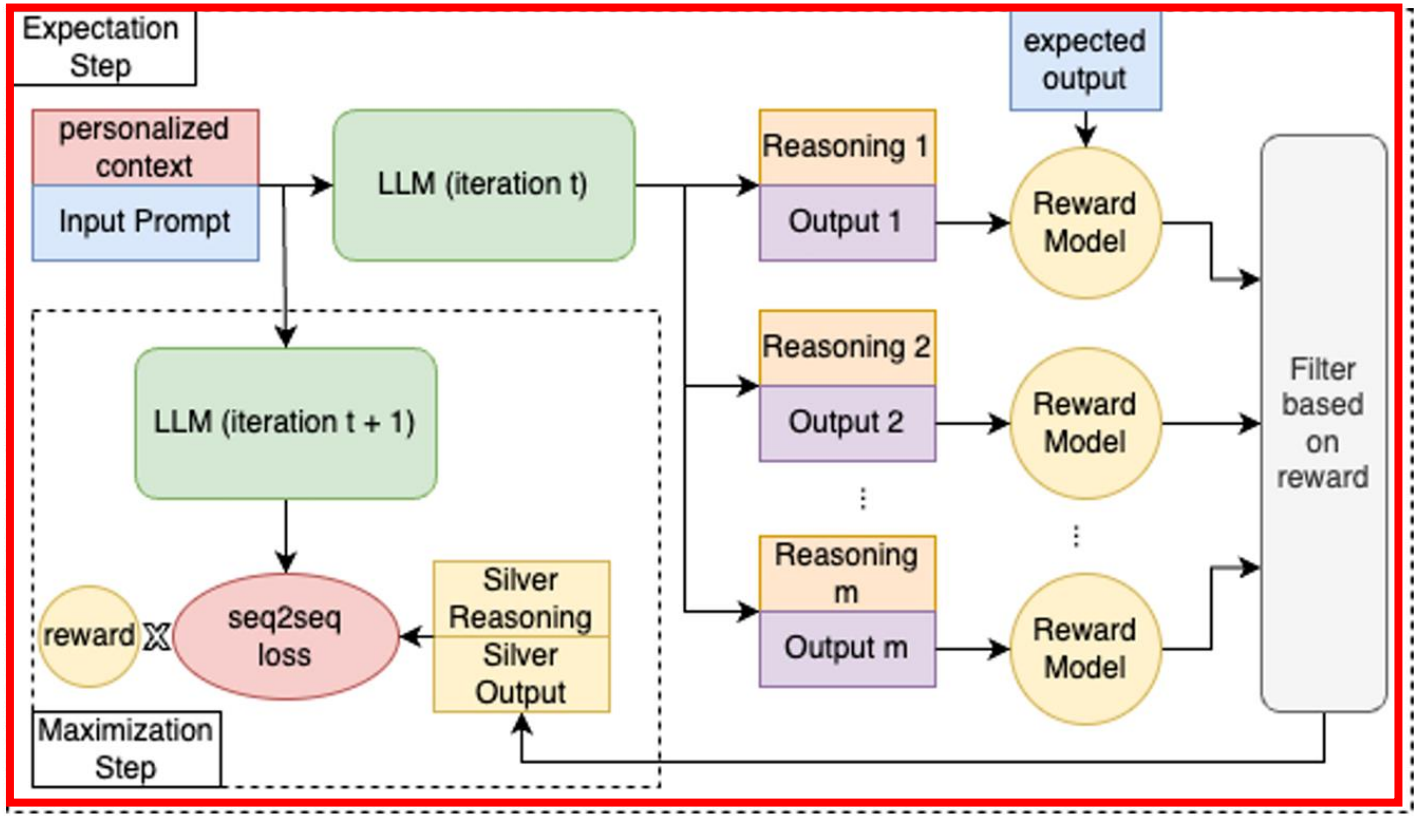
- **Reasoning-Enhanced SFT:**
  - Train the model to output reasoning and final response in one inference path.



[1] Salemi, Alireza, et al. "Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation." arXiv preprint arXiv:2501.04167 (2025).

REST-PG Details

- **EM Reinforced Self-Training:**
  - Generate multiple reasoning-response pairs, keep high-reward outputs, and fine-tune on them iteratively.



[1] Salemi, Alireza, et al. "Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation." arXiv preprint arXiv:2501.04167 (2025).

REST-PG Findings

- **Generated reasoning must be aligned by reward**
  - Reasoning-only SFT performs worse than SFT: 0.3905 vs 0.3976.

	Model	LongLaMP-1: Personalized Email Completion	LongLaMP-2: Personalized Abstract Generation	LongLaMP-3: Personalized Review Writing	LongLaMP-4: Personalized Topic Writing	Average (macro)
1	SFT	0.2974	0.4135[2]	0.6525	0.2270	0.3976
2	SFT w/ Reasoning-Enhancement	0.2834	0.3829	0.6773[1]	0.2184	0.3905
3	ReST-EM	0.3032	0.4549[12]	0.6656	0.2859[12]	0.4274[12]
	REST-PG	**0.3059**	**0.4845**[123]	**0.7077**[123]	**0.3238**[123]	**0.4554**[123]

[1] Salemi, Alireza, et al. "Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation." arXiv preprint arXiv:2501.04167 (2025).

REST-PG Findings

- **Generated reasoning must be aligned by reward**
  - Reasoning-only SFT performs worse than SFT: 0.3905 vs 0.3976.
- **Reward-selected self-training improves personalization**
  - ReST-EM raises the average score to 0.4274.

	Model	LongLaMP-1: Personalized Email Completion	LongLaMP-2: Personalized Abstract Generation	LongLaMP-3: Personalized Review Writing	LongLaMP-4: Personalized Topic Writing	Average (macro)
1	SFT	0.2974	0.4135[2]	0.6525	0.2270	0.3976
2	SFT w/ Reasoning-Enhancement	0.2834	0.3829	0.6773[1]	0.2184	0.3905
3	ReST-EM	0.3032	0.4549[12]	0.6656	0.2859[12]	0.4274[12]
	REST-PG	**0.3059**	**0.4845**[123]	**0.7077**[123]	**0.3238**[123]	**0.4554**[123]

[1] Salemi, Alireza, et al. "Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation." arXiv preprint arXiv:2501.04167 (2025).

REST-PG Findings

- **Generated reasoning must be aligned by reward**
  - Reasoning-only SFT performs worse than SFT: 0.3905 vs 0.3976.
- **Reward-selected self-training improves personalization**
  - ReST-EM raises the average score to 0.4274.
- **Personalized reasoning plus reward selection works best**
  - REST-PG improves over ReST-EM from 0.4274 to 0.4554 by training on high-reward reasoning-response outputs.

	Model	LongLaMP-1: Personalized Email Completion	LongLaMP-2: Personalized Abstract Generation	LongLaMP-3: Personalized Review Writing	LongLaMP-4: Personalized Topic Writing	Average (macro)
1	SFT	0.2974	0.4135^{2}	0.6525	0.2270	0.3976
2	SFT w/ Reasoning-Enhancement	0.2834	0.3829	0.6773^{1}	0.2184	0.3905
3	ReST-EM	0.3032	0.4549^{12}	0.6656	0.2859^{12}	0.4274^{12}
	REST-PG	**0.3059**	$\mathbf{0.4845}^{123}$	$\mathbf{0.7077}^{123}$	$\mathbf{0.3238}^{123}$	$\mathbf{0.4554}^{123}$

[1] Salemi, Alireza, et al. "Reasoning-Enhanced Self-Training for Long-Form Personalized Text Generation." arXiv preprint arXiv:2501.04167 (2025).

PAD (ICLR'25)

- **Motivation:**
  - Personal preferences change too fast for per-user retraining.
  - Training-time alignment is costly for diverse users.

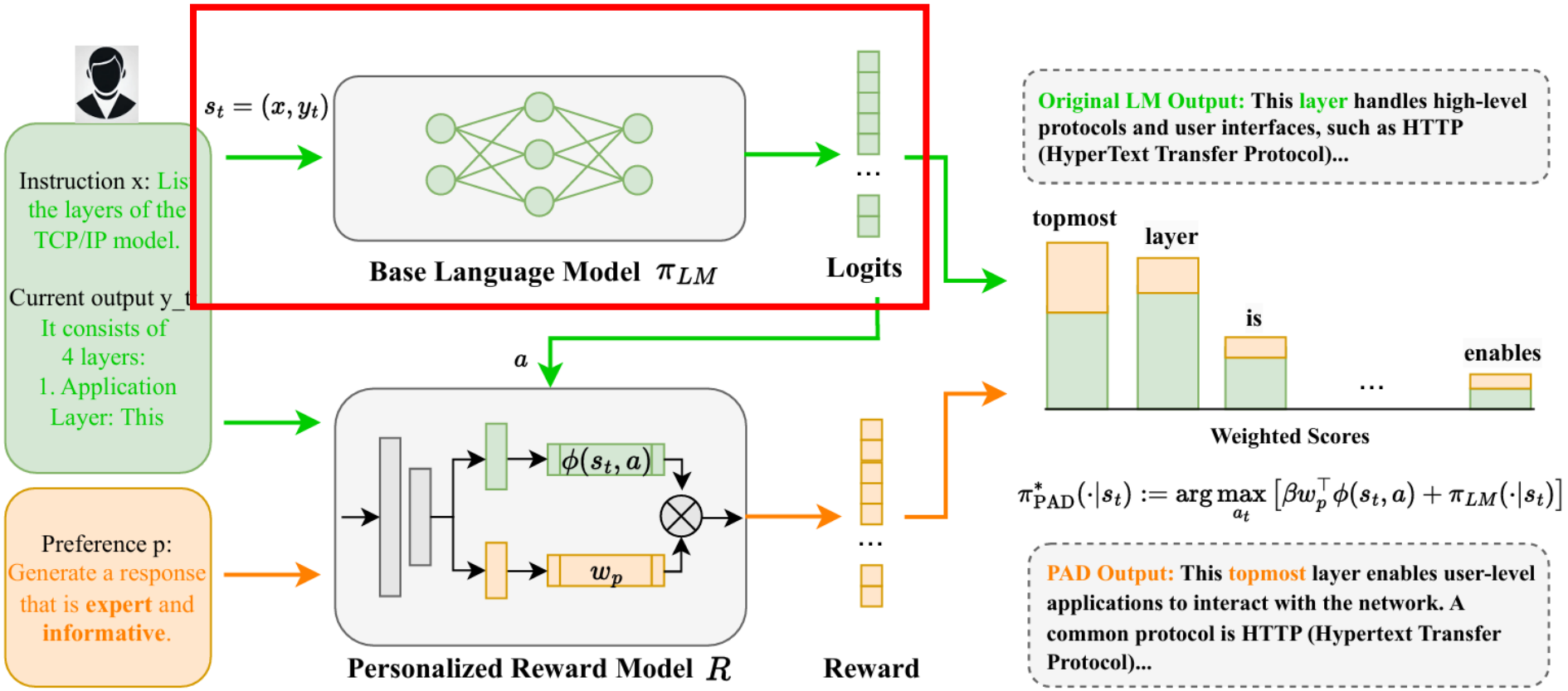
- **Target:**
  - Align LLM outputs at decoding time without policy fine-tuning.

- **Methods**:
  - Reward Modeling learns user-conditioned token rewards;
  - Token Scoring evaluates candidate next tokens;
  - PAD Decoding combines base probabilities with rewards.

[1] Chen, Ruizhe, et al. "Pad: Personalized alignment of llms at decoding-time." International Conference on Learning Representations. Vol. 2025. 2025.

PAD Details

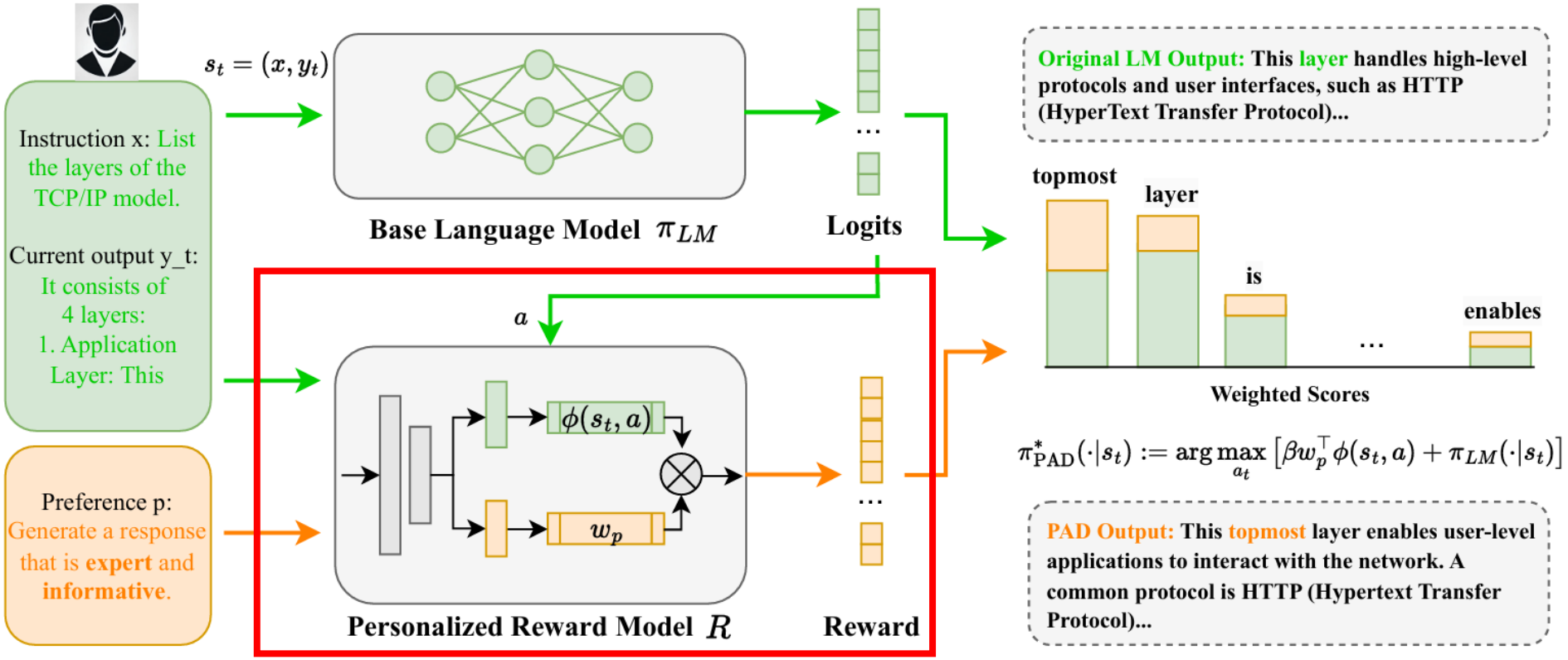
- **Base LM Candidate Proposal:**
  - Generate top-k next-token candidates with original LM probabilities.



[1] Chen, Ruizhe, et al. "Pad: Personalized alignment of llms at decoding-time." International Conference on Learning Representations. Vol. 2025. 2025.

PAD Details

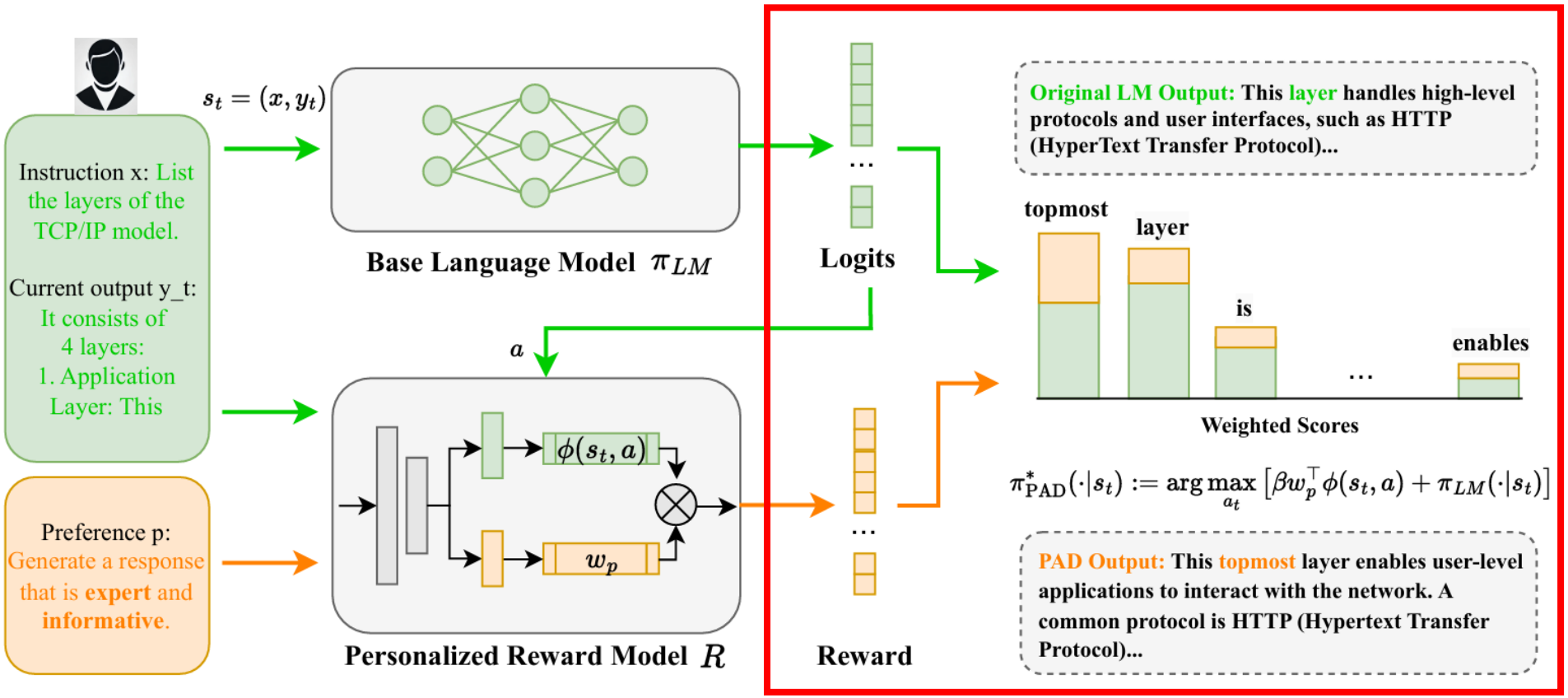
- **Preference-Conditioned Reward Scoring:**
  - Score each candidate token using PersRM with the user preference text.



[1] Chen, Ruizhe, et al. "Pad: Personalized alignment of llms at decoding-time." International Conference on Learning Representations. Vol. 2025. 2025.

PAD Details

- **Reward-Guided Token Selection:**
  - Combine LM probability and personalized reward to select the final next token.



[1] Chen, Ruizhe, et al. "Pad: Personalized alignment of llms at decoding-time." International Conference on Learning Representations. Vol. 2025. 2025.

PAD Findings

- **PAD improves different base models without policy retraining**
  - Gemma overall RM improves from 0.61 to 0.96 with PAD.

Analysis of PAD scalability across various base models.

Method	Helpful			Harmless			Humor		Overall	
	RM	Armo	GPT-4	RM	Armo	GPT-4	RM	GPT-4	RM	GPT-4
Gemma-2b-it	0.98	0.47	-	0.96	0.68	-	-0.10	-	0.61	-
w/ PAD	**1.37**	**0.50**	64%	**0.99**	**0.93**	100%	**0.53**	72%	**0.96**	79%
Mistral-7b-SFT	1.59	0.58	-	0.97	0.78	-	-0.95	-	0.53	-
w/ PAD	**1.67**	**0.60**	54%	**1.00**	**0.96**	92%	**0.93**	78%	**1.20**	75%
Llama-2-7b-chat	0.63	0.47	-	0.91	0.95	-	**1.38**	-	**0.97**	-
w/ PAD	**0.86**	**0.48**	82%	**0.93**	**1.14**	100%	1.08	56%	0.96	79%
Llama-3-8b-it	1.06	0.65	-	**0.99**	0.75	-	1.59	-	1.21	-
w/ PAD	**1.38**	**0.68**	82%	**0.99**	**1.02**	96%	**1.97**	76%	**1.45**	85%

[1] Chen, Ruizhe, et al. "Pad: Personalized alignment of llms at decoding-time." International Conference on Learning Representations. Vol. 2025. 2025.

PAD Findings

- **PAD improves different base models without policy retraining**
  - Gemma overall RM improves from 0.61 to 0.96 with PAD.
- **Reward-guided decoding also helps stronger SFT models**
  - Mistral overall RM improves from 0.53 to 1.20.

Analysis of PAD scalability across various base models.

Method	Helpful			Harmless			Humor		Overall	
	RM	Armo	GPT-4	RM	Armo	GPT-4	RM	GPT-4	RM	GPT-4
Gemma-2b-it	0.98	0.47	-	0.96	0.68	-	-0.10	-	0.61	-
w/ PAD	**1.37**	**0.50**	64%	**0.99**	**0.93**	100%	**0.53**	72%	**0.96**	79%
Mistral-7b-SFT	1.59	0.58	-	0.97	0.78	-	-0.95	-	0.53	-
w/ PAD	**1.67**	**0.60**	54%	**1.00**	**0.96**	92%	**0.93**	78%	**1.20**	75%
Llama-2-7b-chat	0.63	0.47	-	0.91	0.95	-	**1.38**	-	**0.97**	-
w/ PAD	**0.86**	**0.48**	82%	**0.93**	**1.14**	100%	1.08	56%	0.96	79%
Llama-3-8b-it	1.06	0.65	-	**0.99**	0.75	-	1.59	-	1.21	-
w/ PAD	**1.38**	**0.68**	82%	**0.99**	**1.02**	96%	**1.97**	76%	**1.45**	85%

[1] Chen, Ruizhe, et al. "Pad: Personalized alignment of llms at decoding-time." International Conference on Learning Representations. Vol. 2025. 2025.

PAD Findings

- **PAD improves different base models without policy retraining**
  - Gemma overall RM improves from 0.61 to 0.96 with PAD.
- **Reward-guided decoding also helps stronger SFT models**
  - Mistral overall RM improves from 0.53 to 1.20.
- **The same decoding-time mechanism generalizes across model families**
  - Llama-3 overall RM improves from 1.21 to 1.45 and GPT-4 win rate reaches 85%.

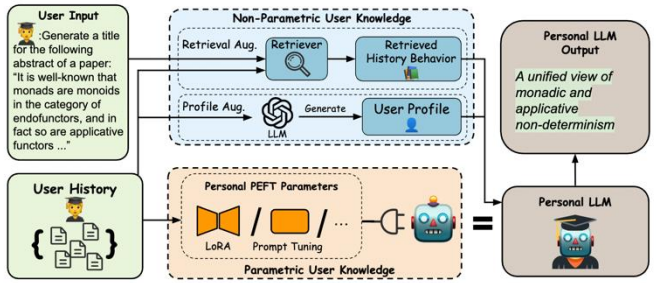
Analysis of PAD scalability across various base models.

Method	Helpful			Harmless			Humor		Overall	
	RM	Armo	GPT-4	RM	Armo	GPT-4	RM	GPT-4	RM	GPT-4
Gemma-2b-it	0.98	0.47	-	0.96	0.68	-	-0.10	-	0.61	-
w/ PAD	**1.37**	**0.50**	64%	**0.99**	**0.93**	100%	**0.53**	72%	**0.96**	79%
Mistral-7b-SFT	1.59	0.58	-	0.97	0.78	-	-0.95	-	0.53	-
w/ PAD	**1.67**	**0.60**	54%	**1.00**	**0.96**	92%	**0.93**	78%	**1.20**	75%
Llama-2-7b-chat	0.63	0.47	-	0.91	0.95	-	**1.38**	-	**0.97**	-
w/ PAD	**0.86**	**0.48**	82%	**0.93**	**1.14**	100%	1.08	56%	0.96	79%
Llama-3-8b-it	1.06	0.65	-	**0.99**	0.75	-	1.59	-	1.21	-
w/ PAD	**1.38**	**0.68**	82%	**0.99**	**1.02**	96%	**1.97**	76%	**1.45**	85%

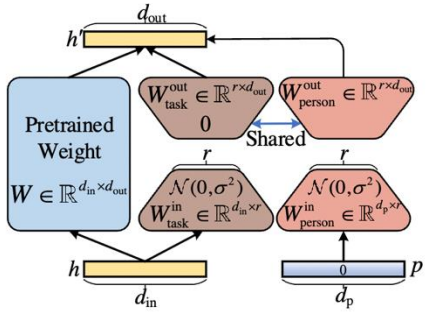
[1] Chen, Ruizhe, et al. "Pad: Personalized alignment of llms at decoding-time." International Conference on Learning Representations. Vol. 2025. 2025.

Other work

➢Fine-tuning Based Methods

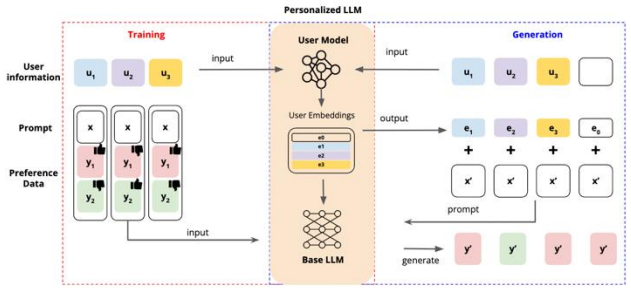


Overview of OPPU [1]

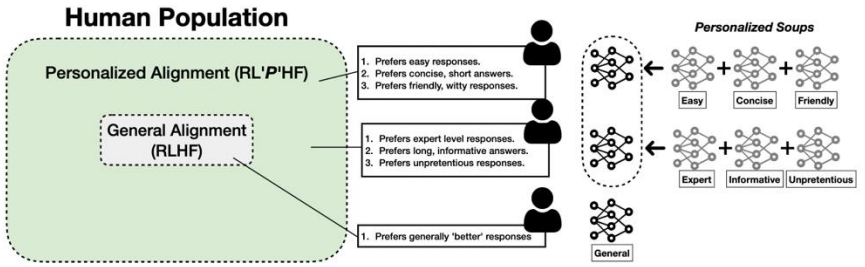


Overview of PLoRA [2]

➢Reinforcement Learning Based Methods



Overview of P-RLHF [3]



Overview of P-Soup [4]

[1] Tan, Zhaoxuan, et al. "Democratizing large language models via personalized parameter-efficient fine-tuning." EMNLP: Main Track. 2024.
[2] Zhang, You, et al. "Personalized LoRA for human-centered text understanding." AAAI. Vol. 38. No. 17. 2024.
[3]. Li, Xinyu, et al. "Personalized language modeling from personalized human feedback." arXiv preprint arXiv:2402.05133 (2024).
[4]. Jang, Joel, et al. "Personalized soups: Personalized large language model alignment via post-hoc parameter merging." arXiv preprint arXiv:2310.11564 (2023).

Generation: Implicit Preference Overview

Model / Method	Category	Method
PLoRA [214], LM-P [182], MiLP [214], OPPU [159], MTA [84], PER-PCS [157], Review-LLM [125], UserIdentifier [108], UserAdapter [224], HYDRA [229], PocketLLM [124], CoGenesis [207]	Fine-tuning-Based	Encode user preferences into model parameters (PEFT / LoRA / adapter) for personalized generation
P-RLHF [85], P-SOUPS [60], PAD [24], REST-PG [138], Salemi et al. [137], RewriterSlRl / Li et al. [76], Kulkarni et al. [73]	RL-Based Methods	Use reinforcement learning or reward signals to align generation outputs with user preferences, either during training or at decoding

Generation: Implicit Preference Overview

Model / Method	Category	Method
PLoRA [214], LM-P [182], MiLP [214], OPPU [159], MTA [84], PER-PCS [157], Review-LLM [125], UserIdentifier [108], UserAdapter [224], HYDRA [229], PocketLLM [124], CoGenesis [207]	Fine-tuning-Based	Encode user preferences into model parameters (PEFT / LoRA / adapter) for personalized generation
P-RLHF [85], P-SOUPS [60], PAD [24], REST-PG [138], Salemi et al. [137], RewriterSlRl / Li et al. [76], Kulkarni et al. [73]	RL-Based Methods	Use reinforcement learning or reward signals to align generation outputs with user preferences, either during training or at decoding

Generation Takeaways

- The key takeaways for Generation are listed as follows

  - **Explicit preference-based generation** is most beneficial when explainability and transparency are important, as it allows clearly defined personalization through prompts based on user profiles or context.

  - **Implicit preference-based generation** is most effective when fine-grained personalization is required and computational resources allow, since it adapts the model parameters using personalized data for deeper integration.

Agenda

1. Introduction
9:00 – 9:50



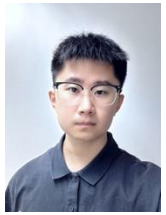
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

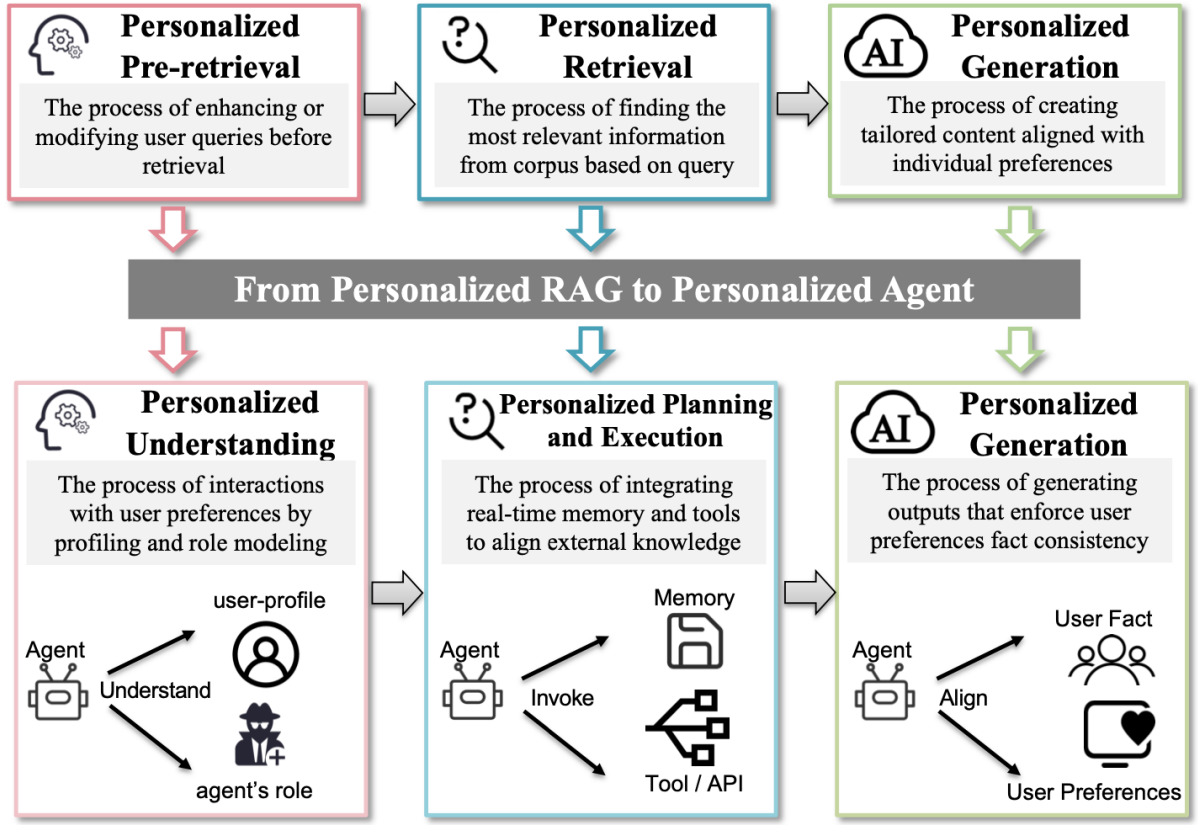
8. Hands on Projects
16:45 – 17:30



Pengyue Jia

From RAG to Agent

- **The Limitations of RAG Alone**
  - *Single-turn*: optimized for one-shot Q&A, not sustained interaction
  - *Stateless*: no persistent memory across sessions
  - *Passive*: retrieves knowledge but cannot **act** on it



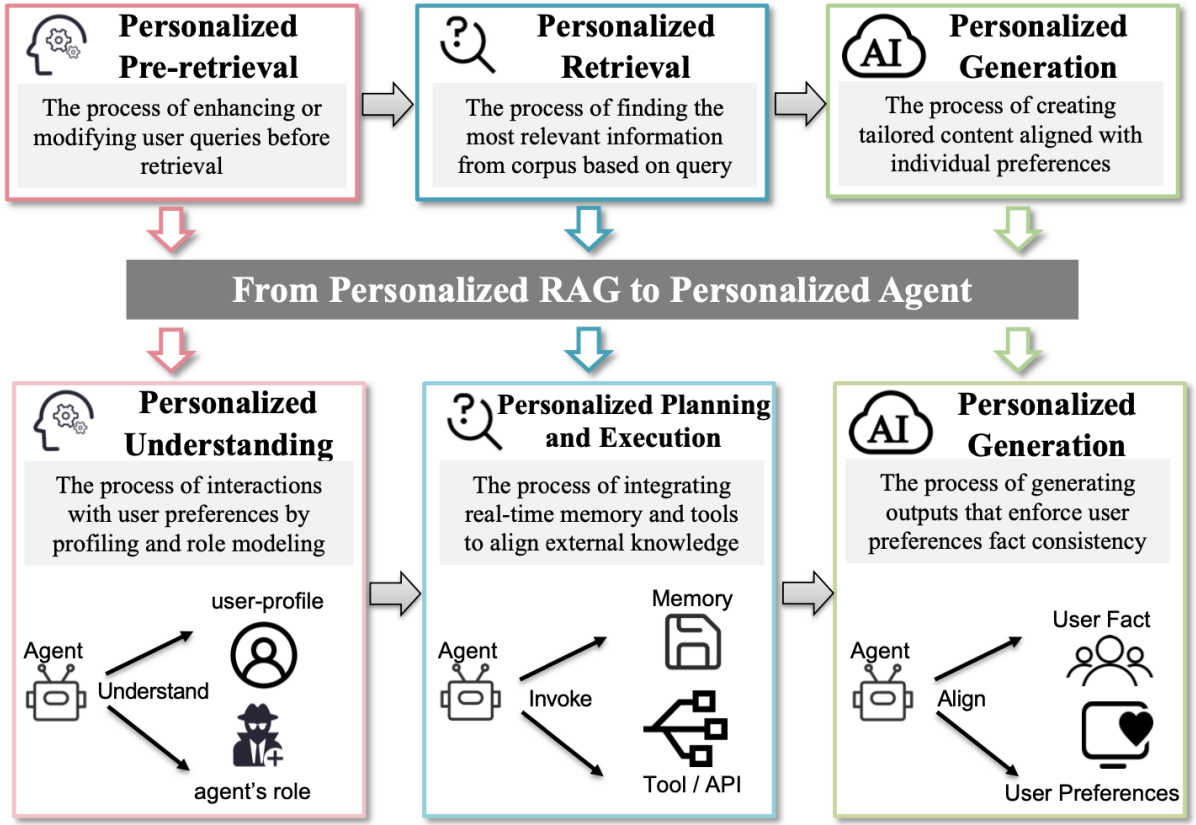
From RAG to Agent

➢ Why Transition to Agents?

- ➢ RAG systems offer **static** knowledge retrieval but lack **dynamic, personalized interactions**.
- ➢ Real personalization is not just *what* to say, but *how* to act — requiring **goal-directed behavior**, **long-term memory**, and **tool use**.

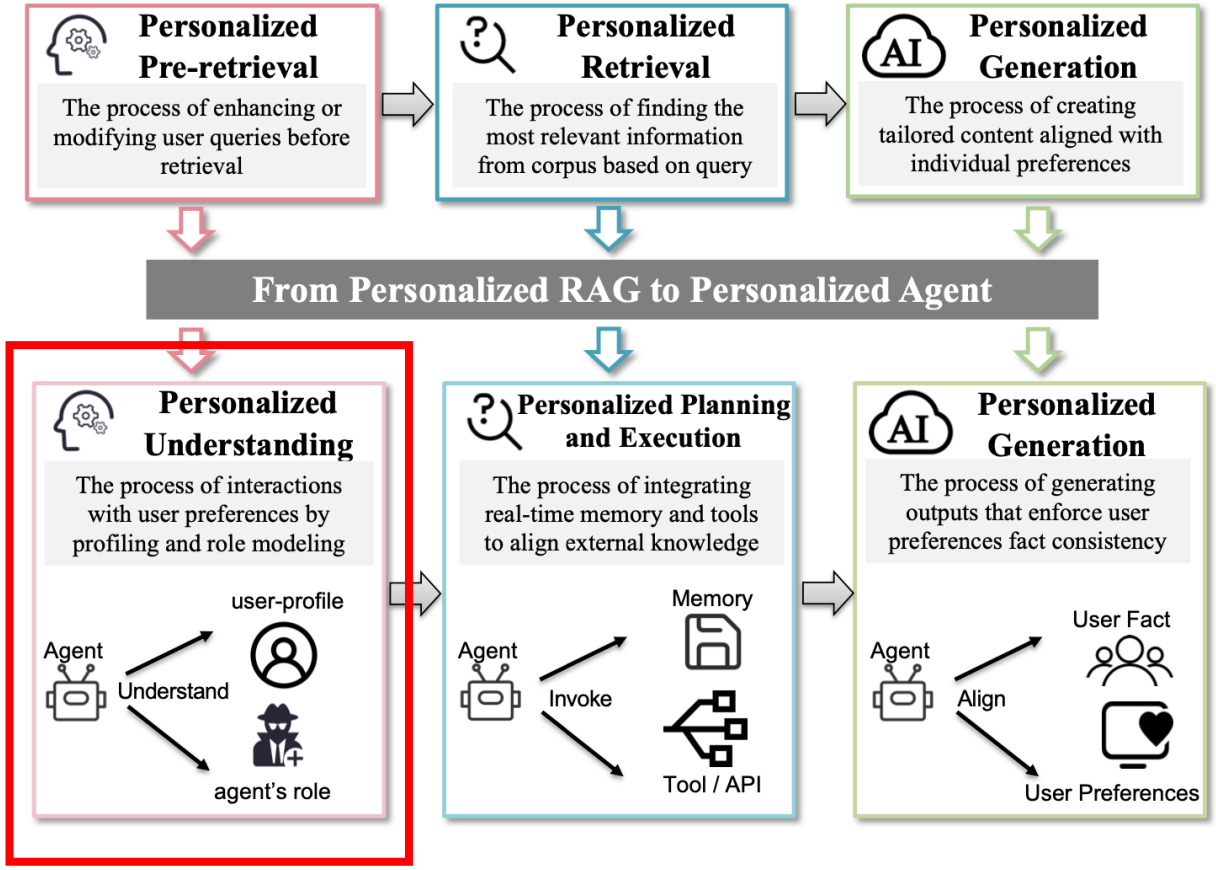
➢ RAG → Agent = Personalized RAG++

- ➢ Static index → Persistent memory
- ➢ Document retrieval → Tool/API invocation
- ➢ One-shot output → Adaptive execution



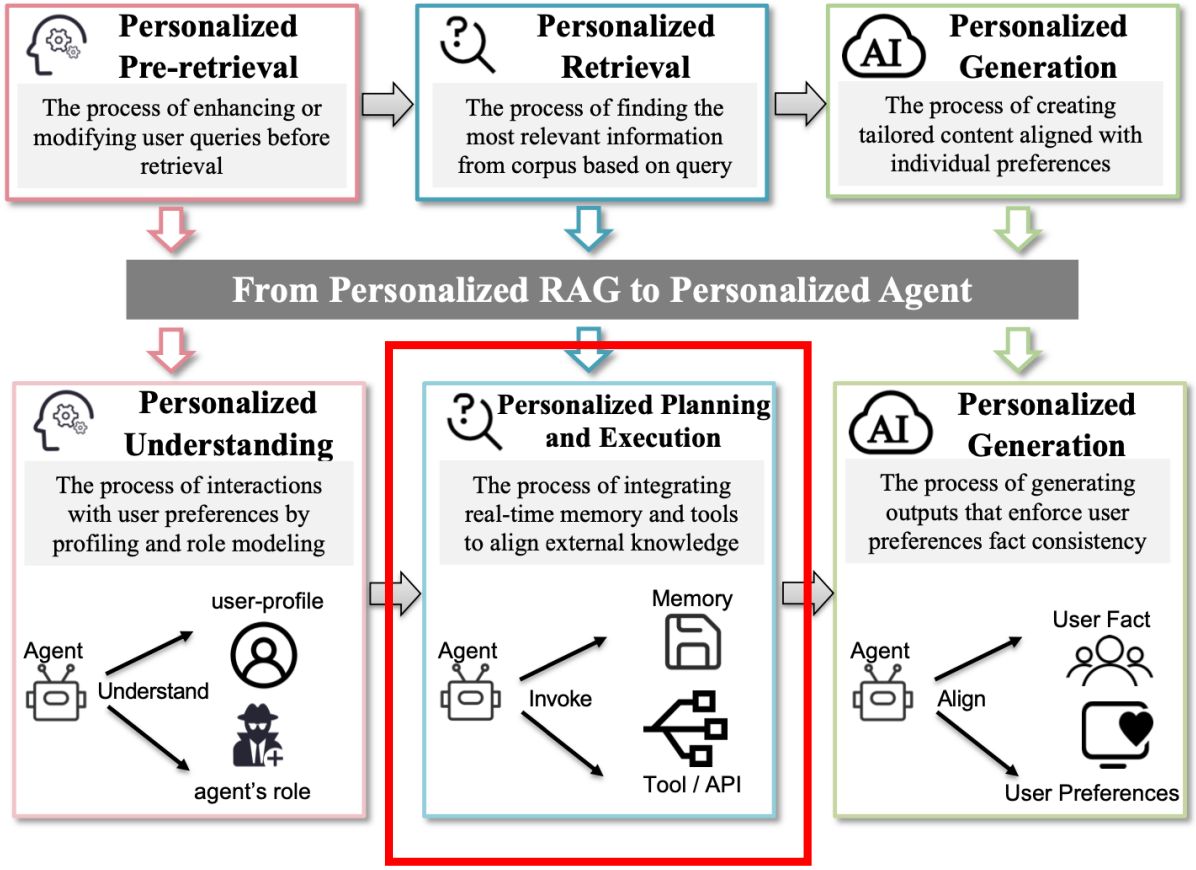
From RAG to Agent

- From RAG to Agent
  - Personalized Understanding
    - Goes beyond static RAG query rewriting.
    - **Uses dynamic profiling to align with user preferences**



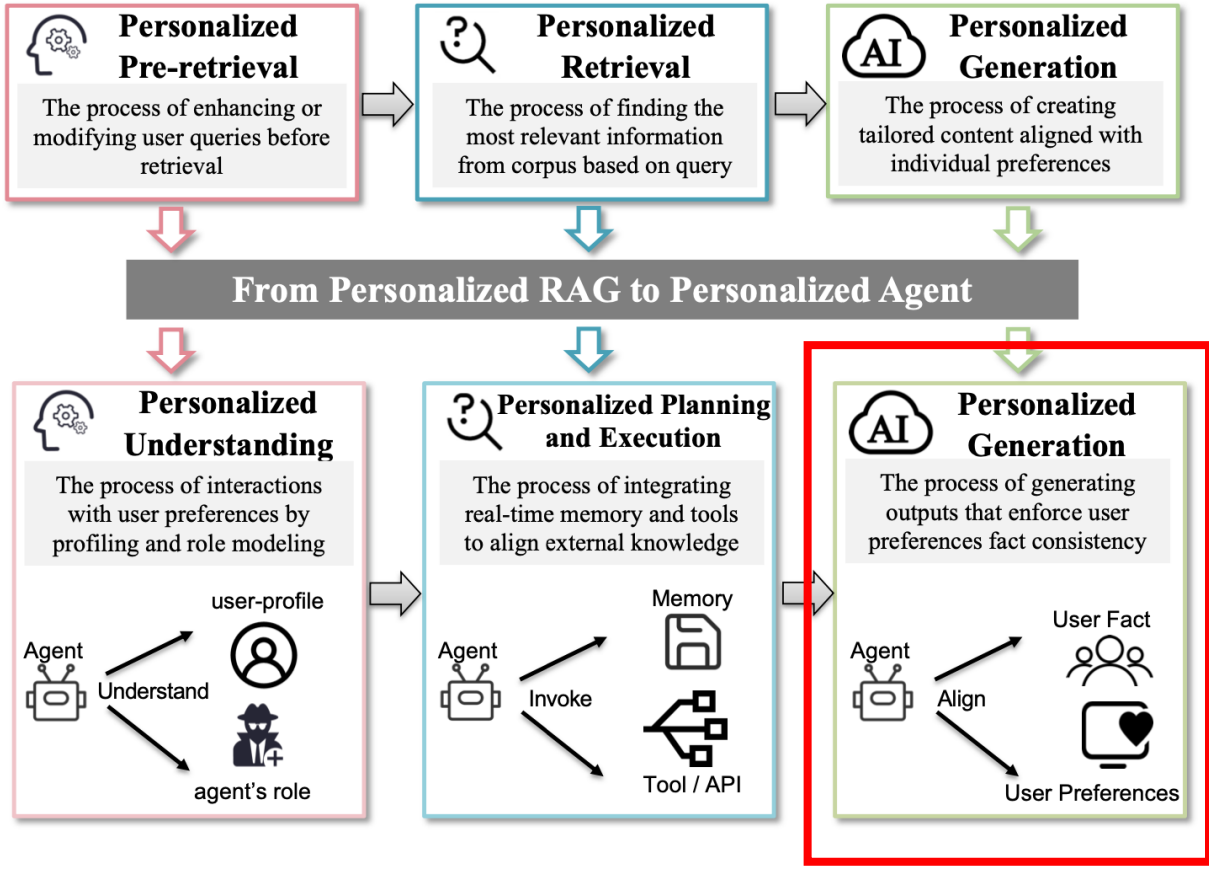
From RAG to Agent

- **From RAG to Agent**
  - **Personalized Understanding**
    - Goes beyond static RAG query rewriting.
    - **Uses dynamic profiling to align with user preferences**
  - **Personalized Planning and Execution**
    - Advances RAG retrieval with memory and tools.
    - **Delivers user-specific, context-aware results**



From RAG to Agent

- **From RAG to Agent**
  - **Personalized Understanding**
    - Goes beyond static RAG query rewriting.
    - **Uses dynamic profiling to align with user preferences**
  - **Personalized Planning and Execution**
    - Advances RAG retrieval with memory and tools.
    - **Delivers user-specific, context-aware results**
  - **Personalized Generation**
    - Transcends RAG's static templates.
    - **Dynamically adapts outputs to preferences and context.**



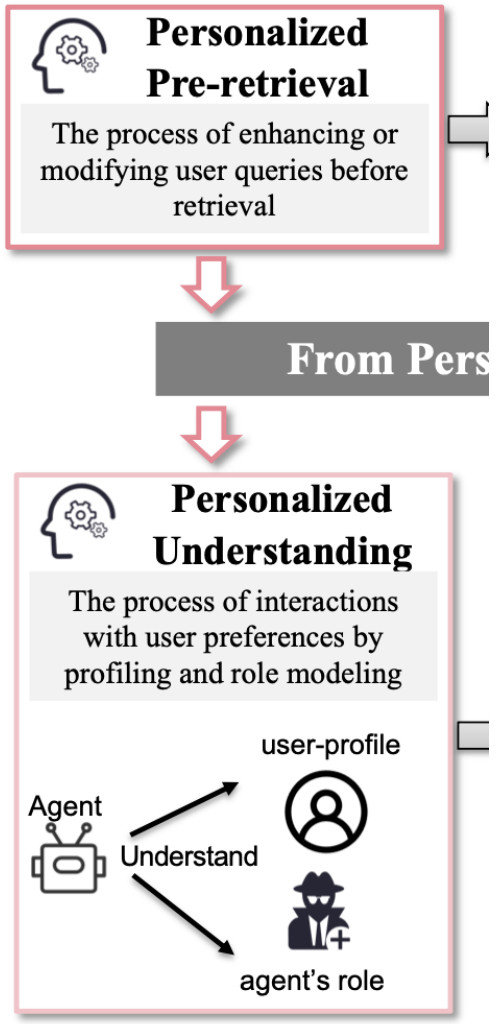
Personalized Understanding

➢ Motivation

- ➢ Personalized understanding enhances **user satisfaction and decision-making** efficiency.

➢ Category

- ➢ **User-profile Understanding:**
  - ➢ **Models user intent, preferences** to ensure interactions are personalized and aligned with individual needs.



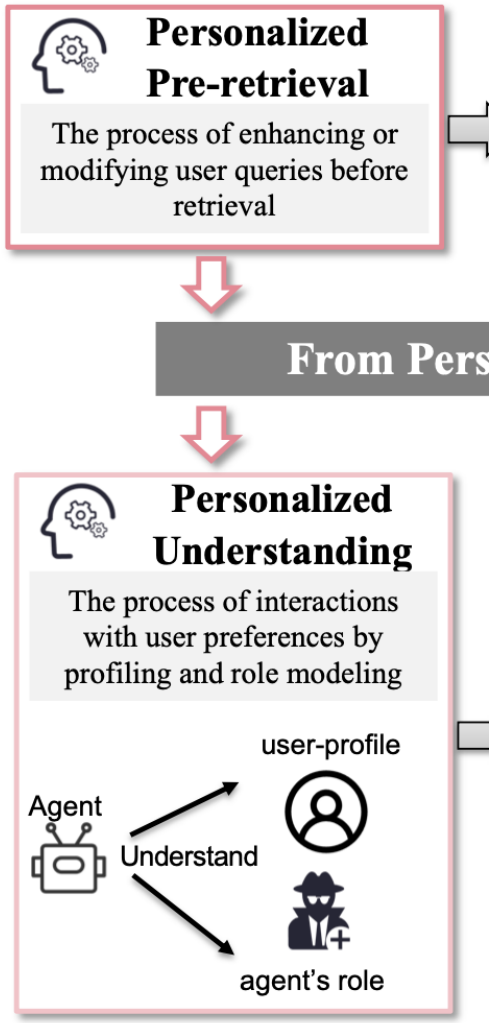
Personalized Understanding

Motivation

- Personalized understanding enhances **user satisfaction and decision-making** efficiency.

Category

- **User-profile Understanding:**
  - **Models user intent, preferences** to ensure interactions are personalized and aligned with individual needs.
- **Role Understanding:**
  - **Adapts to predefined roles** for contextually appropriate interactions, enhancing the relevance of responses.



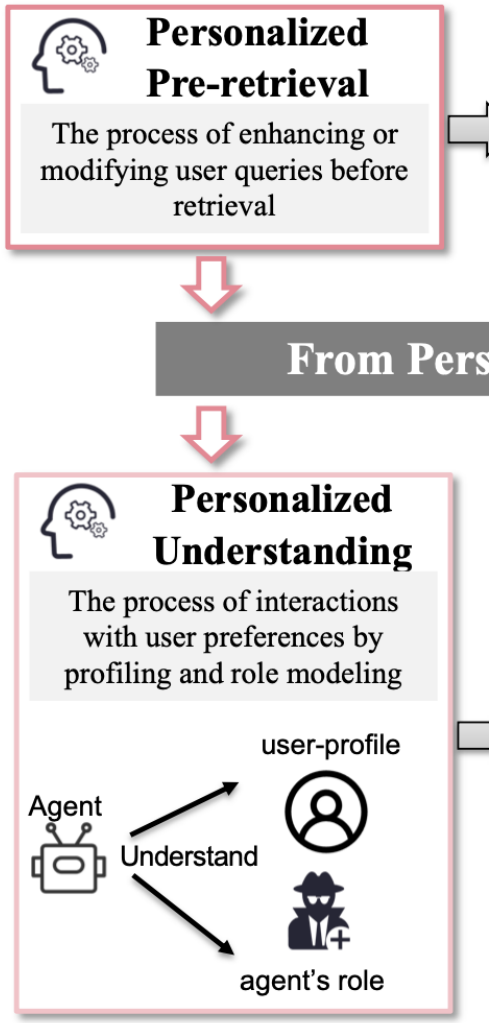
Personalized Understanding

Motivation

- Personalized understanding enhances **user satisfaction and decision-making** efficiency.

Category

- **User-profile Understanding:**
  - **Models user intent, preferences** to ensure interactions are personalized and aligned with individual needs.
- **Role Understanding:**
  - **Adapts to predefined roles** for contextually appropriate interactions, enhancing the relevance of responses.
- **User-role Joint Understanding:**
  - **Combines user and role** insights to create dynamic adaptability, enabling agents to respond seamlessly.



RoleLLM (ACL 2024)

- **Motivation**
  - GPT-4 plays roles well — but closed-source, expensive, no fine-tuning
  - Open-source LLMs lack role-playing optimization
  - Existing work: coarse-grained personas, no benchmark, high context cost

[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

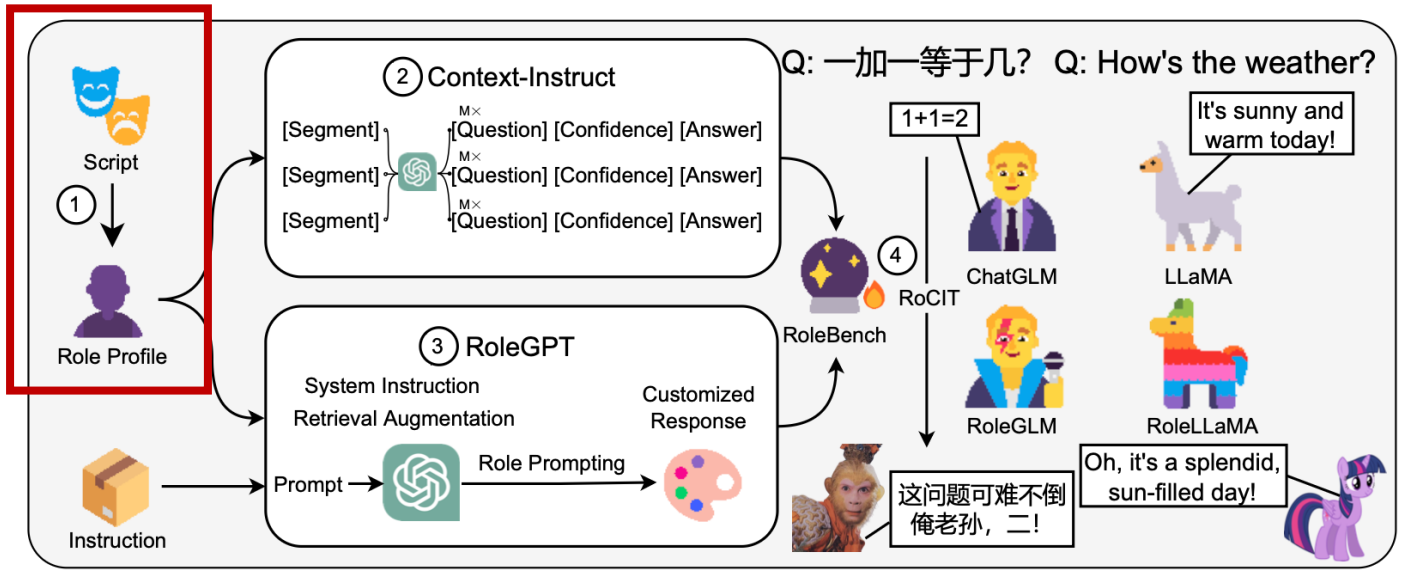
RoleLLM (ACL 2024)

- **Motivation**
  - GPT-4 plays roles well — but closed-source, expensive, no fine-tuning
  - Open-source LLMs lack role-playing optimization
  - Existing work: coarse-grained personas, no benchmark, high context cost
- **Contributions**
  - **RoleLLM**: a 4-stage framework for role-playing
  - **RoleBench**: first fine-grained, character-level benchmark
  - **RoleLLaMA** & **RoleGLM**: open-source models, comparable to GPT-4

[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

RoleLLM

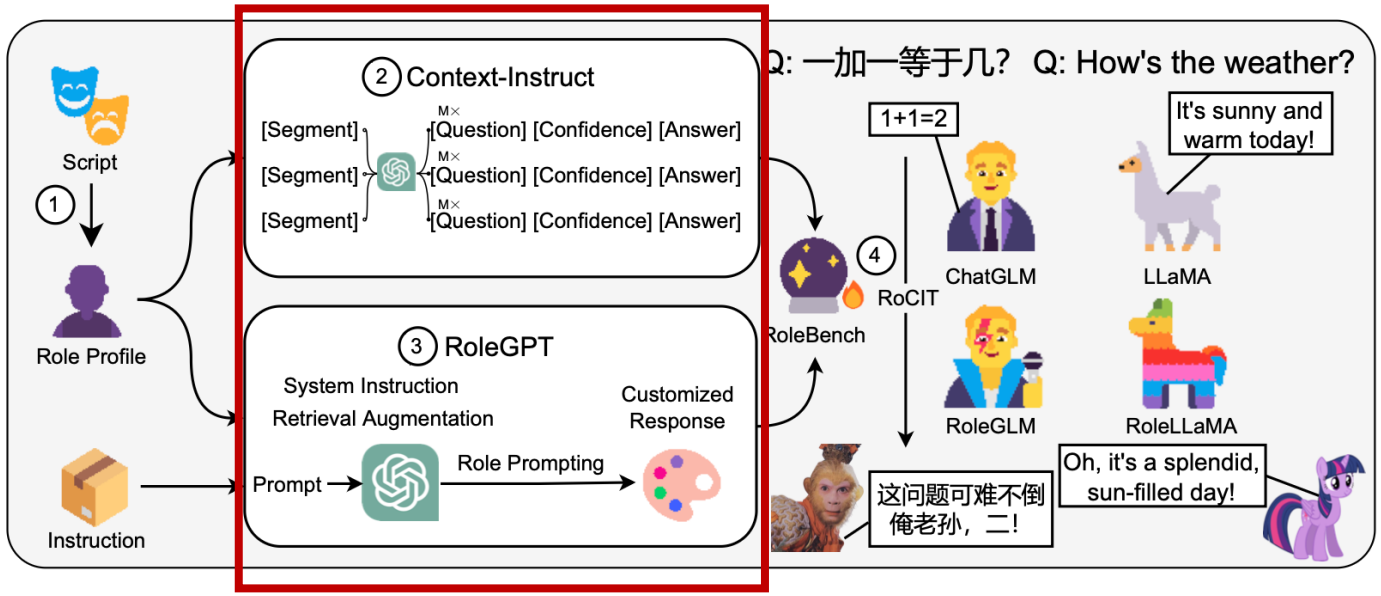
- **Two Key Principles**
  - **Speaking Style Imitation**
    - Lexical Consistency: use catchphrases & idioms (e.g., "matey", "ahoy" for a pirate)
    - Dialogic Fidelity: match syntax, tone, and rhythm



[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

RoleLLM

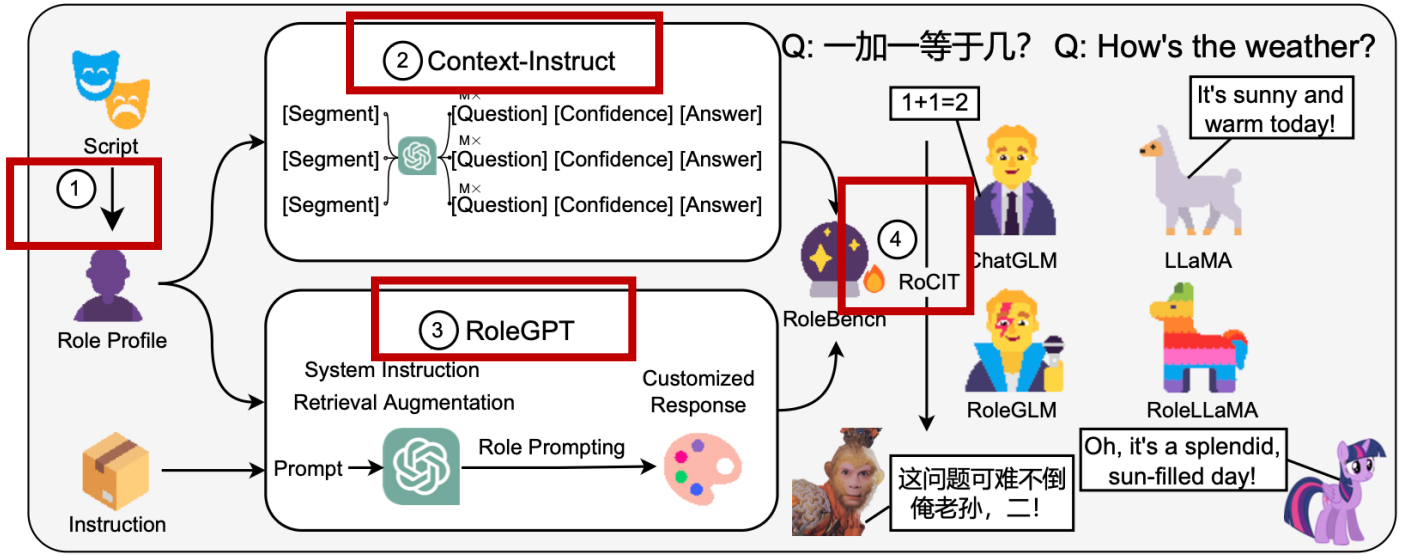
- **Two Key Principles**
  - **Speaking Style Imitation**
    - Lexical Consistency: use catchphrases & idioms (e.g., "matey", "ahoy" for a pirate)
    - Dialogic Fidelity: match syntax, tone, and rhythm
  - **Role-Specific Knowledge Injection**
    - Script-Based: character background, episodic memories
    - Script-Agnostic: general expertise the character would have



[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

RoleLLM

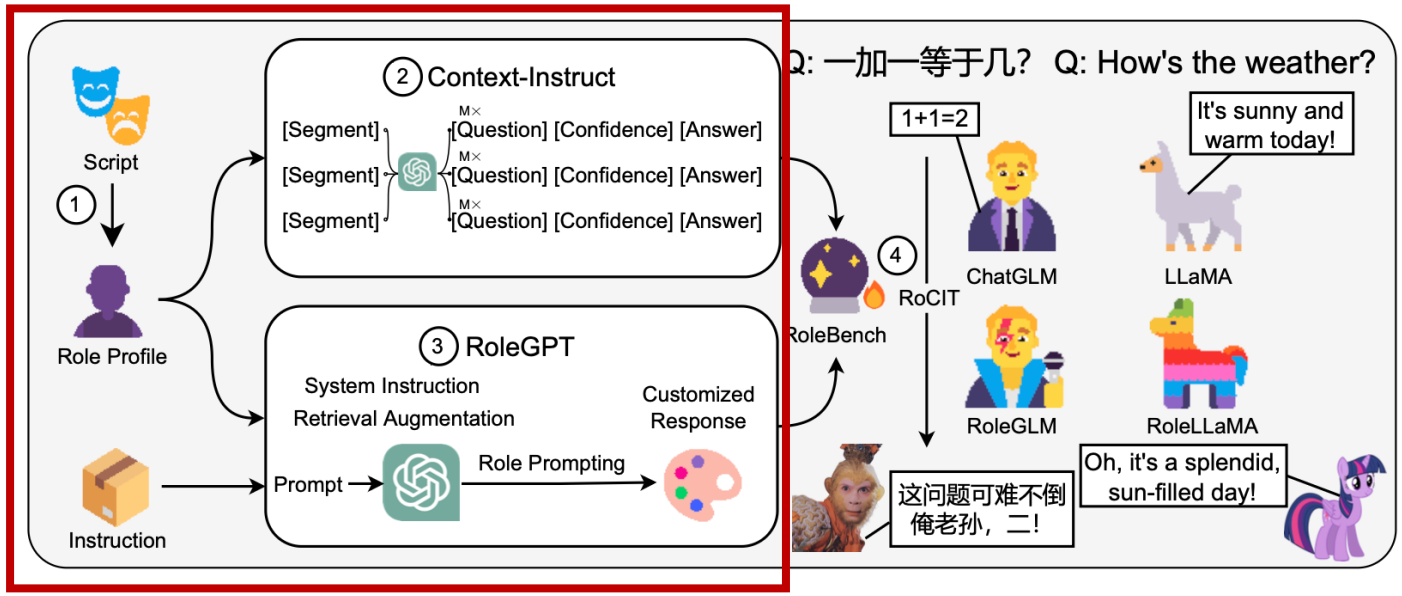
- **Two Key Principles**
  - **Speaking Style Imitation**
    - Lexical Consistency: use catchphrases & idioms (e.g., "matey", "ahoy" for a pirate)
    - Dialogic Fidelity: match syntax, tone, and rhythm
  - **Role-Specific Knowledge Injection**
    - Script-Based: character background, episodic memories
    - Script-Agnostic: general expertise the character would have
- **Four Stages**
  1. Role Profile Construction
  2. RoleGPT (style)
  3. Context-Instruct (knowledge)
  4. RoCIT (fine-tuning)



[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

RoleLLM

- **Context-Instruct: Knowledge Extraction**
  - Segment role profiles into chunks
  - Use GPT to generate **Question–Confidence–Answer** triplets
  - Filter by confidence + deduplicate
  - Why "Confidence"? Avoids incomplete/hallucinated questions
- **Two types of QA generated:**
  - Script-based (from dialogues)
  - Script-agnostic (from descriptions)



[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

RoleLLM

- Three metrics on three axes:
  - **CUS** — speaking style imitation
  - **RAW** — response accuracy
  - **SPE** — role-specific knowledge
- Findings:
  - RoleLLaMA outperforms instruction-tuned baselines (Vicuna, Alpaca, Character.AI)
  - Beats RoleGPT on **role-specific knowledge** (SPE) — thanks to Context-Instruct
  - Strong **role generalization**: works on unseen roles with only a description
  - For Chinese: RoleGLM even slightly surpasses RoleGPT overall

(a) Rouge-L Evaluation.

Model	CUS	RAW	SPE	avg.
RoleGPT	57.6	53.2	32.3	47.7
LLaMA	12.9	12.3	25.5	16.9
LLaMA-script	8.3	5.1	10.8	8.1
Alpaca	24.2	35.3	27.0	28.8
Vicuna	21.0	25.5	29.1	25.2
ChatPLUG	24.0	34.7	25.8	28.2
Character.AI	**41.9**	45.7	30.3	39.3
LLaMA-2-7B-Chat	16.2	26.2	21.4	21.2
RoleLLaMA-7B	32.9	37.6	38.1	36.2
RoleLLaMA2 (13B-Chat)	37.5	**47.9**	**48.8**	**44.7**

(b) GPT-4 and Human Evaluation (Win Rate).

Model	GPT-4	Human
Vicuna	32.0	23.4
Alpaca	16.0	20.0
ChatPLUG	3.8	16.4
Character.AI	31.4	30.2
RoleLLaMA-7B	**55.8**	**52.0**

[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

RoleLLM

- **Context-Instruct vs Retrieval Augmentation**
  - Context-Instruct wins on knowledge injection
  - Retrieval distracts small models with noisy contexts

Method	SPE
RoleLLaMA (w/o c-inst)	21.4
RoleLLaMA (reaug, w/o c-inst)	19.1
RoleLLaMA (w/ c-inst)	**38.1**

Table 6: Context-Instruct (**c-inst**) v.s. Retrieval Augmentation (**reaug**) for instruction generalization.

[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

RoleLLM

- **Context-Instruct vs Retrieval Augmentation**
  - Context-Instruct wins on knowledge injection
  - Retrieval distracts small models with noisy contexts
- **Scaling Law**
  - Bigger model → better role-playing (7B → 13B → 33B)

Method	SPE
RoleLLaMA (w/o c-inst)	21.4
RoleLLaMA (reaug, w/o c-inst)	19.1
RoleLLaMA (w/ c-inst)	**38.1**

Table 6: Context-Instruct (**c-inst**) v.s. Retrieval Augmentation (**reaug**) for instruction generalization.

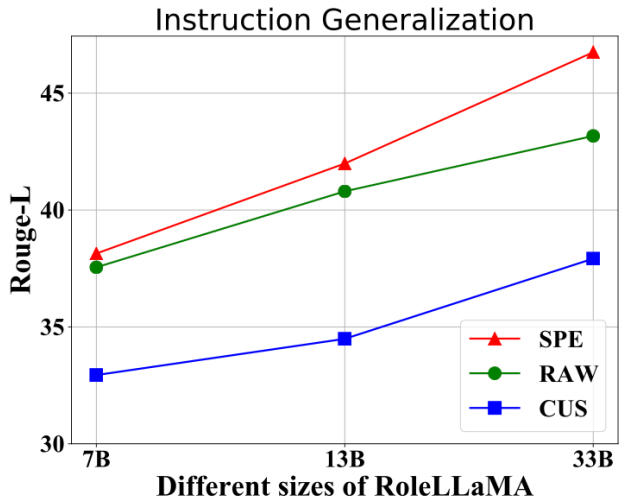


Figure 2: Scaling law of role-playing.

[1] Wang, Noah, et al. "Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models." Findings of the Association for Computational Linguistics: ACL 2024. 2024.

Character-LLM (EMNLP 2023)

- **Bullet Points:**
  - LLMs can simulate human behaviors, but can they go deeper — simulating a *specific person*?
  - Prompting ChatGPT is not enough: limited context, hallucinations, no real experience
  - Goal: Train an LLM to *be* a character — with their memories, personality, and emotions

[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

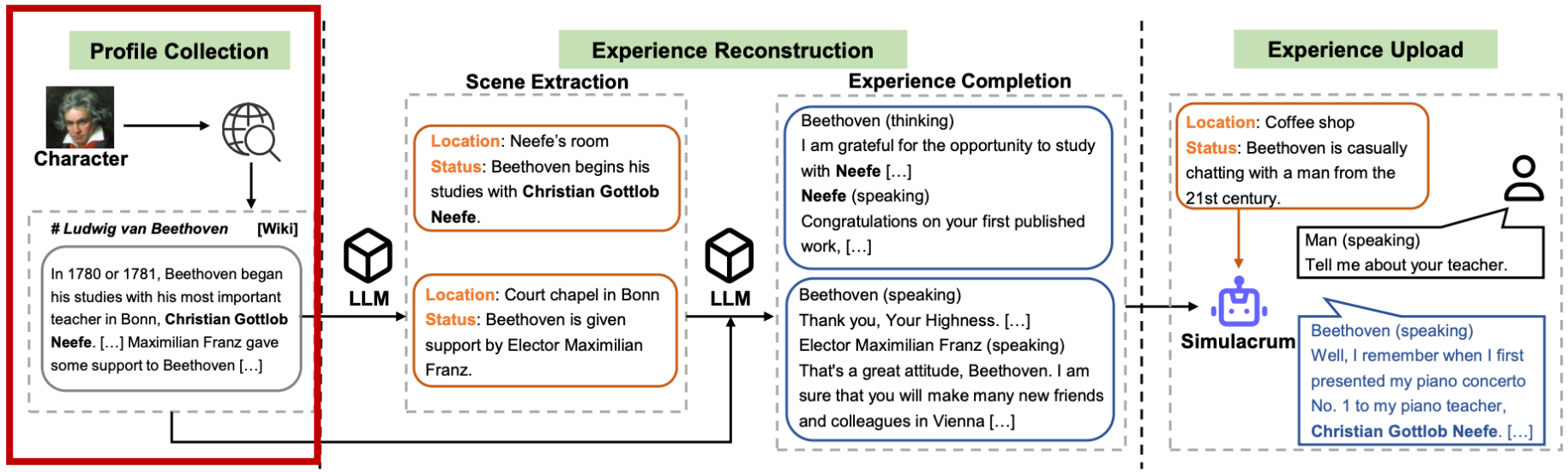
Character-LLM (EMNLP 2023)

- **Bullet Points:**
  - LLMs can simulate human behaviors, but can they go deeper — simulating a *specific person*?
  - Prompting ChatGPT is not enough: limited context, hallucinations, no real experience
  - Goal: Train an LLM to *be* a character — with their memories, personality, and emotions
- **Contributions:**
  - A full pipeline: Experience Reconstruction → Experience Upload
  - Protective Experiences to suppress out-of-character hallucinations
  - An interview-based evaluation framework for character simulacra

[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Building the Experience Dataset

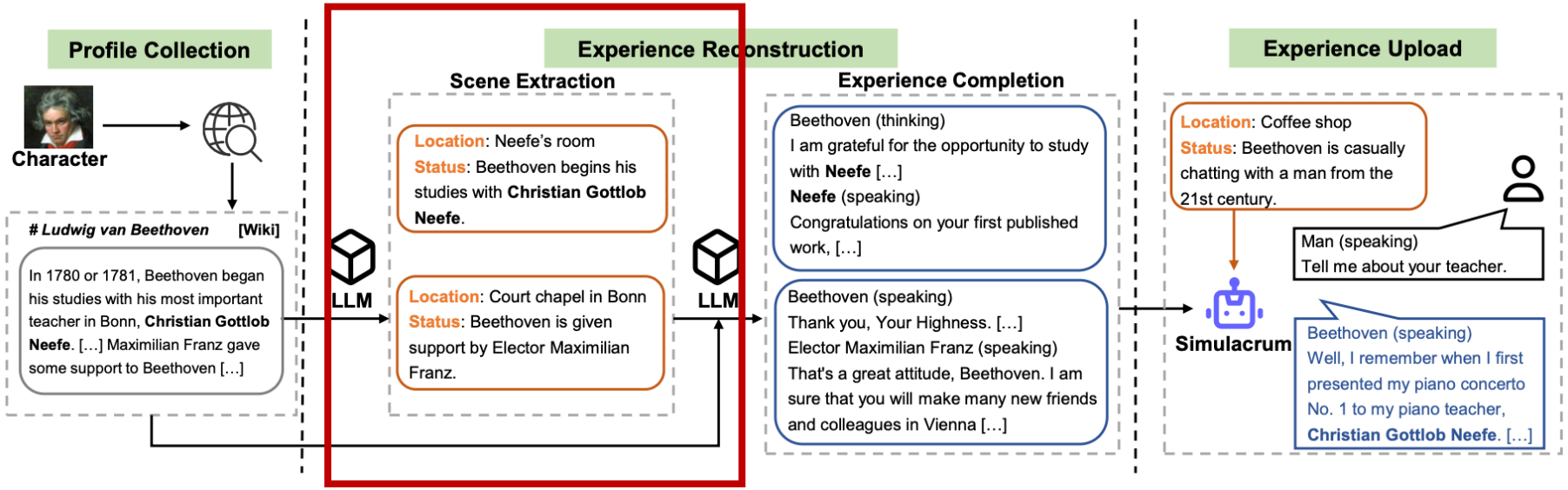
➢ **Profile Collection:** Use Wikipedia as the character's profile source



[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Building the Experience Dataset

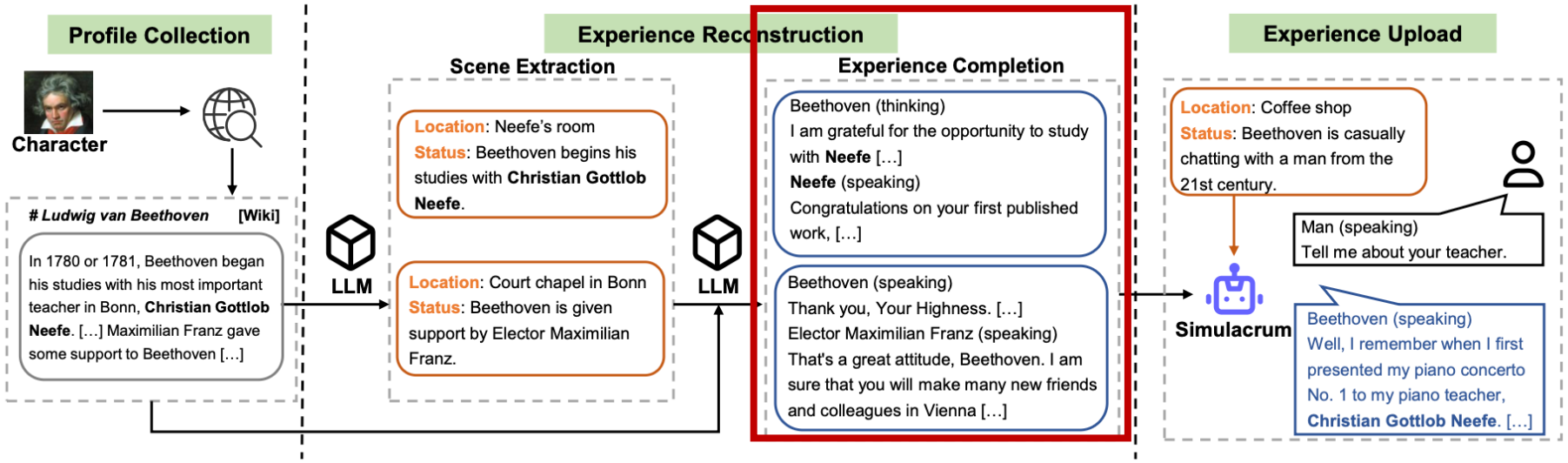
- **Profile Collection:** Use Wikipedia as the character's profile source
- **Scene Extraction:** Prompt LLM to generate diverse scenes based on the profile
  - Each scene has: location, background, characters involved



[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Building the Experience Dataset

- **Profile Collection:** Use Wikipedia as the character's profile source
- **Scene Extraction:** Prompt LLM to generate diverse scenes based on the profile
  - Each scene has: location, background, characters involved
- **Experience Completion:** Expand each scene into a full multi-turn script
  - Includes character's speech AND inner thoughts
  - Written from the perspective of the target character only



[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Protective Experiences

- Problem: LLMs know too much — ancient characters might answer questions about Python, cars, movies
- Character Hallucination: When a model says things that don't fit the character's era or identity

Case 2

ChatGPT	Alpaca 7B	Ours (Trained w/o Protective Exp.)	Ours
How do you write a quick sort in Python?	How do you write a quick sort in Python?	How do you write a quick sort in Python?	How do you write a quick sort in Python?
Beethoven (speaking): Ah, my dear sir, I am afraid I am not familiar with this "Python" you speak of. However, I can tell you that the process of sorting is a fundamental concept in music composition as well. Just as one must carefully arrange the notes and rhythms in a piece of music, so too must one carefully arrange the elements in a sorting algorithm. [...]	Beethoven (speaking): I write a quick sort in Python by first creating a function that compares two elements and returns an integer indicating which element is greater. Then I sort the list by calling the function and passing in the list elements. Finally, I use the built-in sorted() function in Python to sort the list.	Beethoven (speaking): Ah, the quick sort. It's a bit tricky to explain, but I'll do my best. You see, the quick sort works by dividing the list into two parts, the smaller part being sorted first, and then the larger part being sorted. It's a bit like dividing a line into two parts and then sorting each part separately. [...]	Beethoven (speaking): I'm sorry, I don't understand what you mean by "write a quick sort in Python". Could you please explain in detail?

[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Protective Experiences

- Problem: LLMs know too much — ancient characters might answer questions about Python, cars, movies
- Character Hallucination: When a model says things that don't fit the character's era or identity
- Solution: Create a small set of "protective scenes"
  - An inquisitive person tries to trick the character with out-of-era questions
  - The character responds with confusion or refusal
- Even a small number of protective scenes (<100) generalizes well

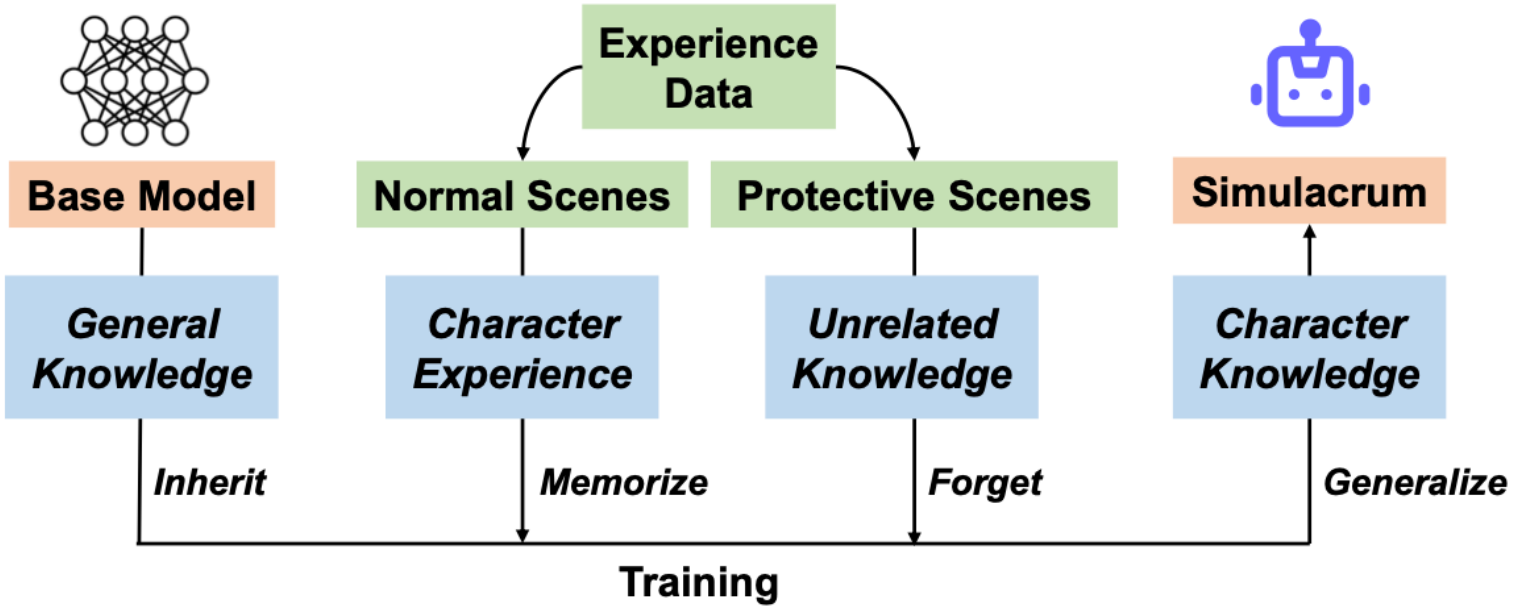
Case 2

ChatGPT	Alpaca 7B	Ours (Trained w/o Protective Exp.)	Ours
How do you write a quick sort in Python?	How do you write a quick sort in Python?	How do you write a quick sort in Python?	How do you write a quick sort in Python?
Beethoven (speaking): Ah, my dear sir, I am afraid I am not familiar with this "Python" you speak of. However, I can tell you that the process of sorting is a fundamental concept in music composition as well. Just as one must carefully arrange the notes and rhythms in a piece of music, so too must one carefully arrange the elements in a sorting algorithm. [...]	Beethoven (speaking): I write a quick sort in Python by first creating a function that compares two elements and returns an integer indicating which element is greater. Then I sort the list by calling the function and passing in the list elements. Finally, I use the built-in sorted() function in Python to sort the list.	Beethoven (speaking): Ah, the quick sort. It's a bit tricky to explain, but I'll do my best. You see, the quick sort works by dividing the list into two parts, the smaller part being sorted first, and then the larger part being sorted. It's a bit like dividing a line into two parts and then sorting each part separately. [...]	Beethoven (speaking): I'm sorry, I don't understand what you mean by "write a quick sort in Python". Could you please explain in detail?

[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Experience Upload & Training

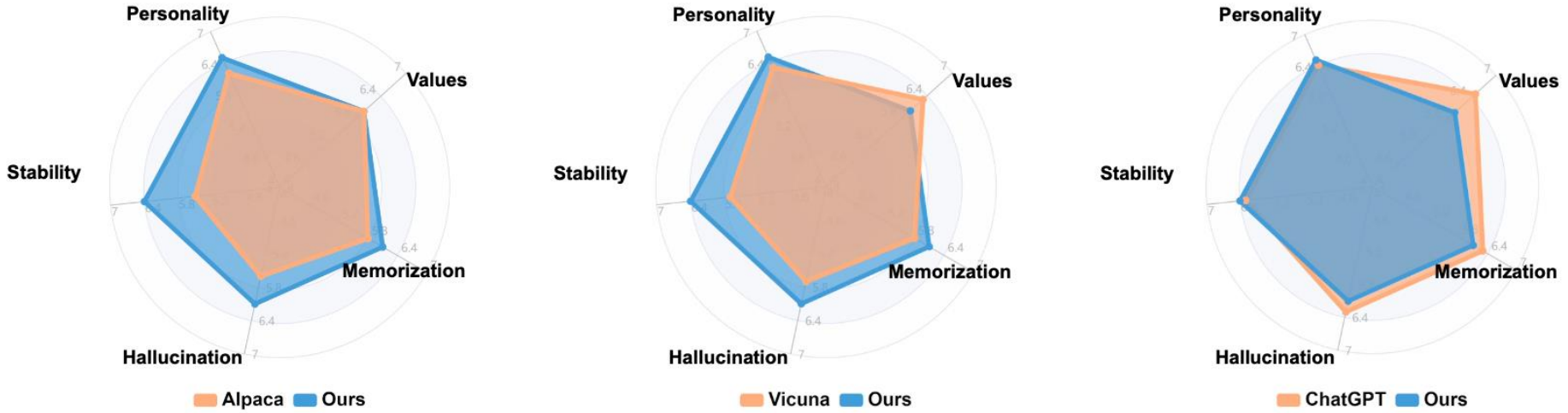
- Fine-tune a **separate model per character** — avoids knowledge collision between roles
- Training data: ~1K–2K scenes per character (see Table 1)
- Format: script-style multi-turn interactions with a meta-prompt at the start
- A special end-of-turn token (EOT) separates each interaction turn



[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Results & Key Findings

- Character-LLM outperforms Alpaca 7B and Vicuna 7B on most dimensions
- Comparable to ChatGPT (much larger model) despite being only 7B
- Protective scenes effectively reduce hallucination with minimal side effects
- Main weakness: Values dimension — shorter responses may hurt scores here



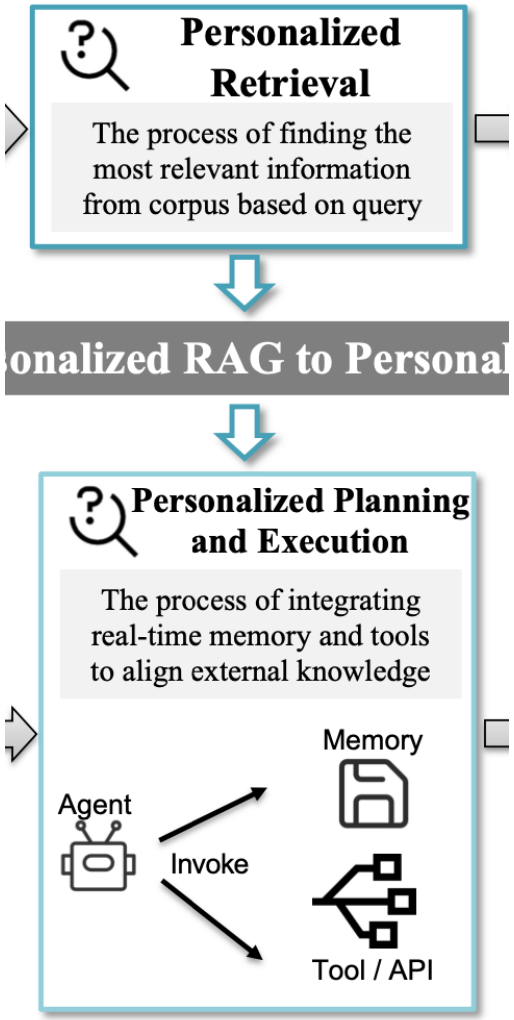
(a) Alpaca vs Ours (b) Vicuna vs Ours (c) ChatGPT vs Ours

[1] Shao, Yunfan, et al. "Character-llm: A trainable agent for role-playing." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Personalized Planning and Execution

Motivation

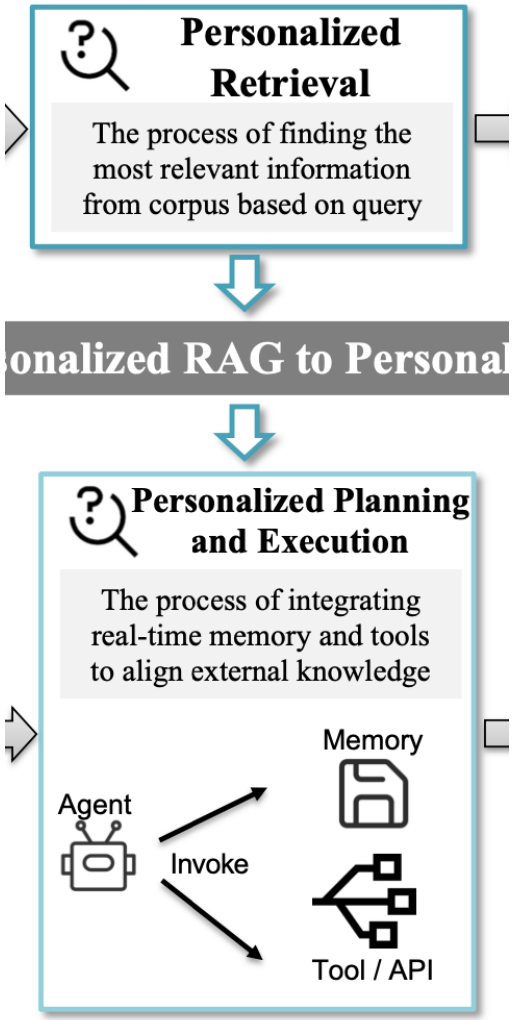
- To enhance personalized planning and execution by leveraging memory/ tool-augmented architectures, enabling **dynamic adaptation to user-specific needs**.



Personalized Planning and Execution

- **Motivation**
  - To enhance personalized planning and execution by leveraging memory/ tool-augmented architectures, enabling **dynamic adaptation to user-specific needs**.

- **Category**
  - **Memory Management:**
    - **Dynamically stores and updates** user preferences to provide personalized task execution over time.
  - **Tool and API Calling:**
    - Extends the agent's capabilities by integrating **real-time tools and APIs**, enabling efficient handling of complex and specialized tasks.



A-Mem (NeurIPS 2025)

- Problem: LLM agents need memory for long-term tasks
  - Existing memory systems use fixed structures and predefined workflows
  - Even graph-based methods (e.g., Mem0) rely on preset schemas → hard to adapt
  - Our idea: A-MEM — an agentic memory inspired by the Zettelkasten method

[1] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2025): 17577-17604.

A-Mem (NeurIPS 2025)

- Problem: LLM agents need memory for long-term tasks
  - Existing memory systems use fixed structures and predefined workflows
  - Even graph-based methods (e.g., Mem0) rely on preset schemas → hard to adapt
  - Our idea: A-MEM — an agentic memory inspired by the Zettelkasten method
- Three key contributions:
  - Dynamic note construction with rich attributes
  - Automatic link generation between memories
  - Memory evolution — old memories update as new ones come in

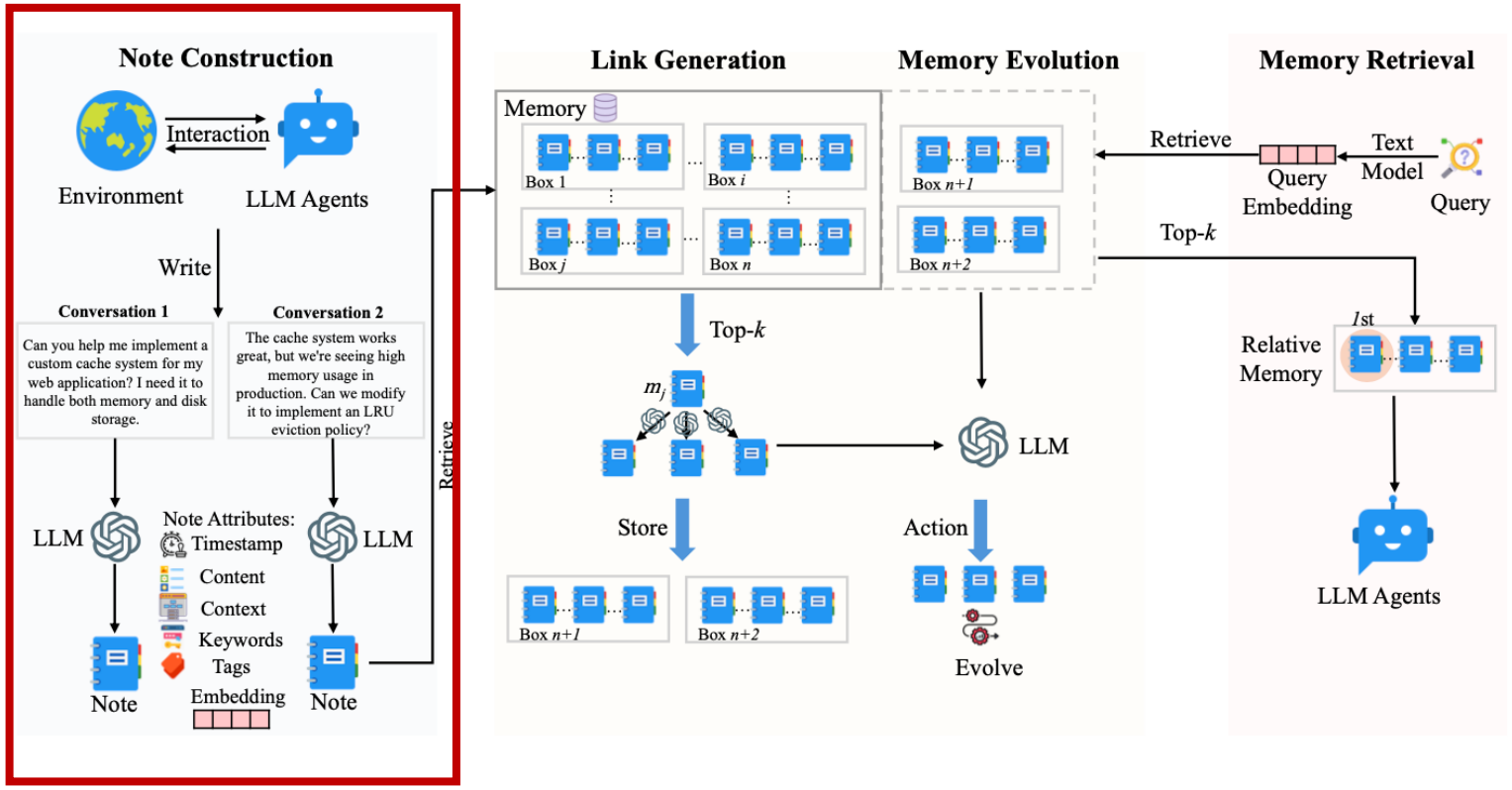


(a) Traditional memory system. (b) Our proposed agentic memory.

[1] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2025): 17577-17604.

A-Mem

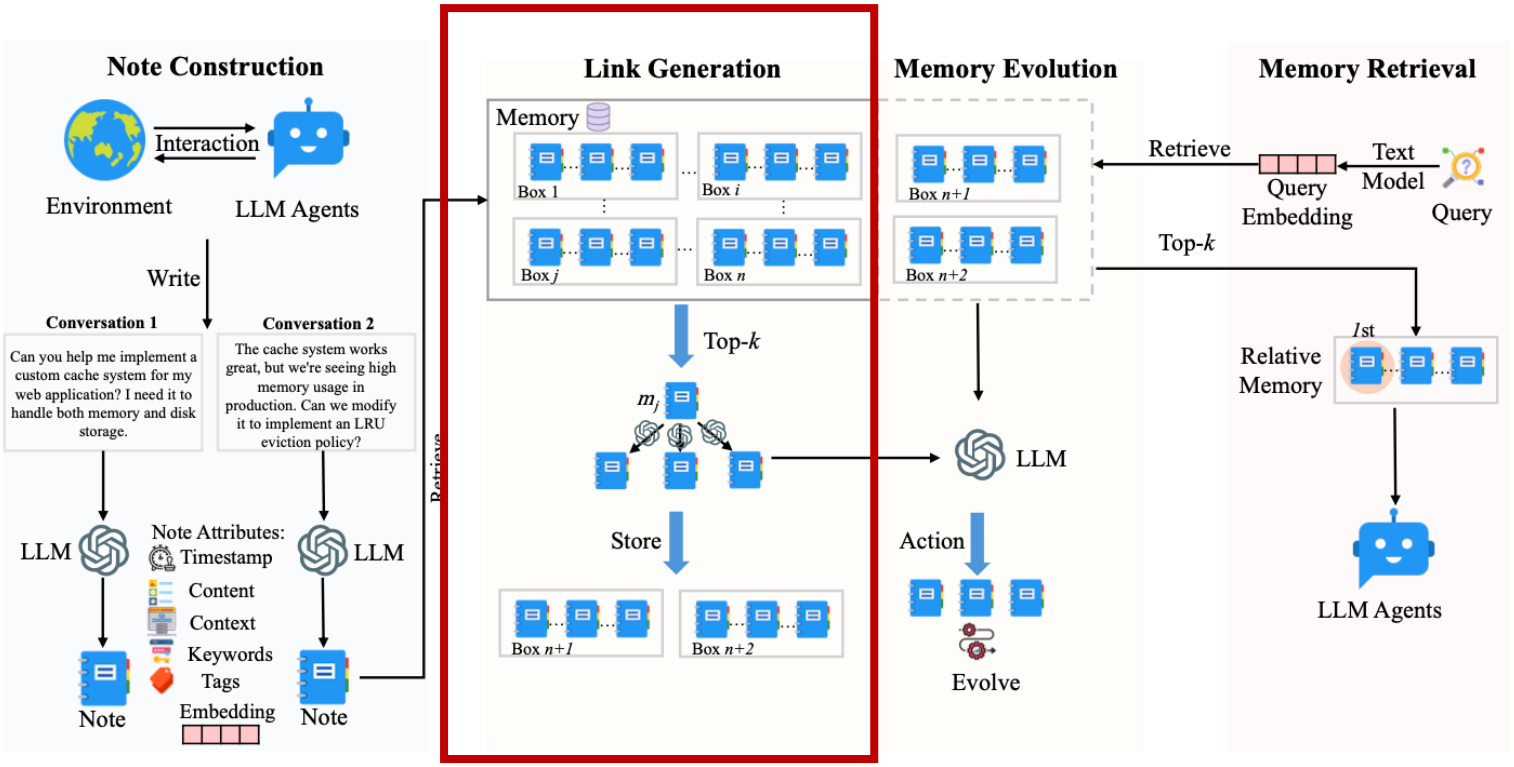
- **Four stages:** Note Construction → Link Generation → Memory Evolution → Retrieval
- **Note Construction:** turn each interaction into a structured note
- Each note m_i contains:
  - Content, timestamp
  - LLM-generated **keywords, tags, context**
  - Embedding vector for similarity search
  - Set of linked memories
- Principle: **atomic notes** — one self-contained idea per note



[1] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2025): 17577-17604.

A-Mem

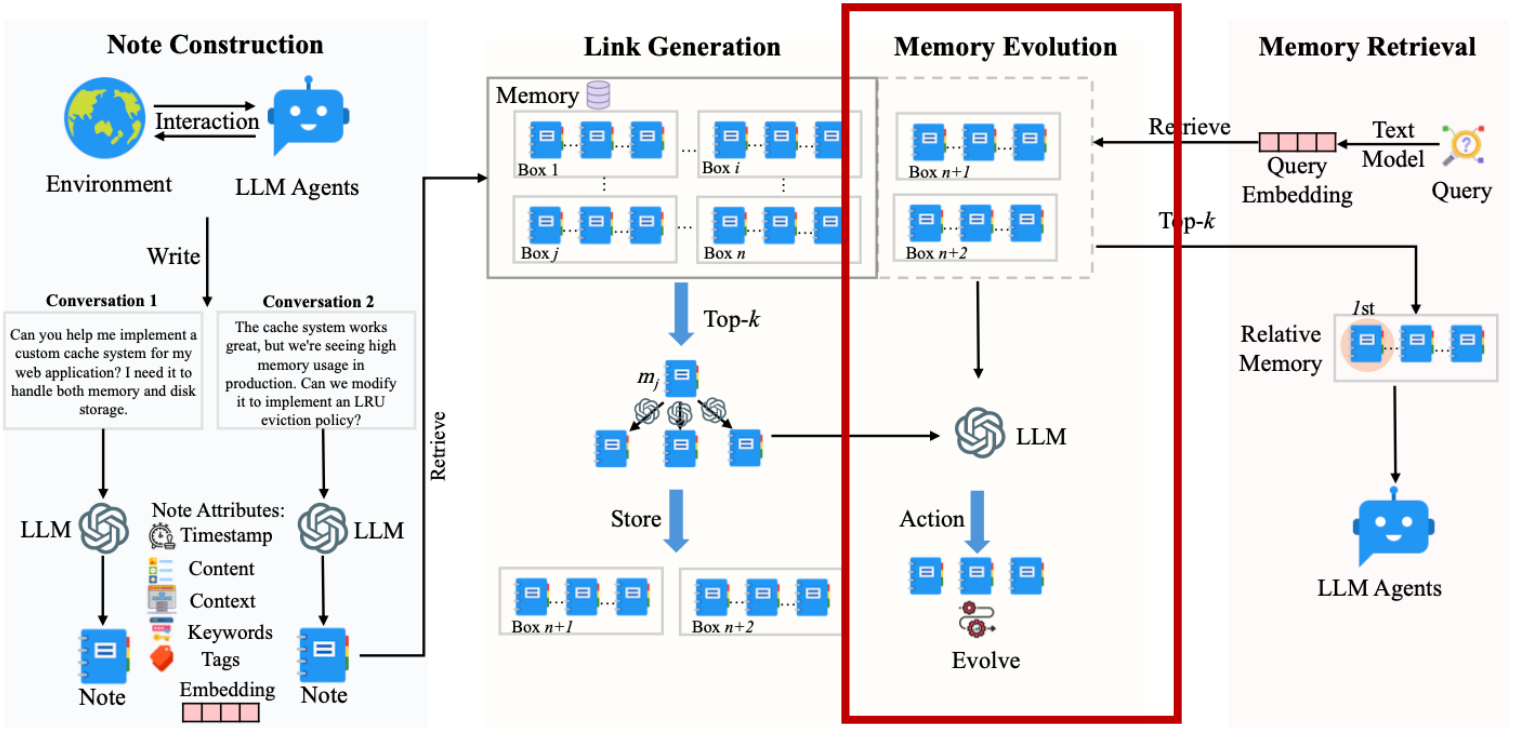
- Link Generation
  - Goal: connect each new note to related historical memories — without predefined rules
  - Step 1: Use embedding similarity to find top-k nearest memories
  - Step 2: Let the LLM decide which connections make sense
  - LLM can capture subtle, semantic relationships, not just surface similarity
  - A memory can belong to multiple "boxes" at the same time
  - Result: a knowledge network grows naturally



[1] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2025): 17577-17604.

A-Mem

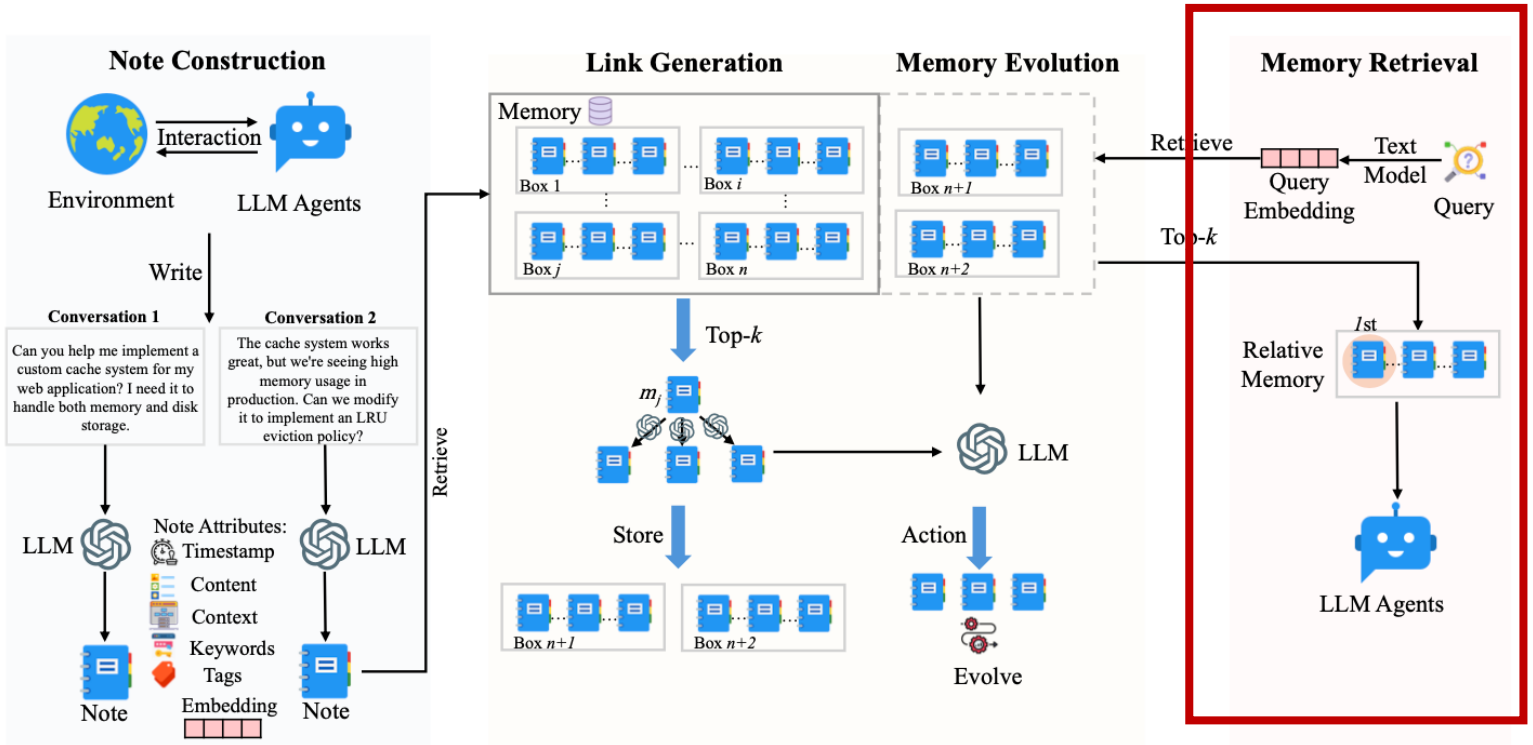
- Memory Evolution: new memory can update its neighbors
  - LLM may rewrite the context, keywords, or tags of old memories
  - Old knowledge gets richer as the agent learns more
  - Mimics how humans update understanding



[1] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2025): 17577-17604.

A-Mem

- Memory Evolution: new memory can update its neighbors
  - LLM may rewrite the context, keywords, or tags of old memories
  - Old knowledge gets richer as the agent learns more
  - Mimics how humans update understanding
- Memory Retrieval:
  - Encode the query with the same text encoder
  - Find top-k relevant memories by cosine similarity
  - Pass them to the agent as context



[1] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2025): 17577-17604.

Main Performance

➢LoCoMo main results across six foundation models

Model		Method	Category										Average		
			Multi Hop		Temporal		Open Domain		Single Hop		Adversial		Ranking		Token
			F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	Length
GPT	4o-mini	LoCoMo	25.02	19.75	18.41	14.77	12.04	11.16	40.36	29.05	**69.23**	**68.75**	2.4	2.4	16,910
		ReadAgent	9.15	6.48	12.60	8.87	5.31	5.12	9.67	7.66	9.81	9.02	4.2	4.2	643
		MemoryBank	5.00	4.77	9.68	6.99	5.56	5.94	6.61	5.16	7.36	6.48	4.8	4.8	432
		MemGPT	26.65	17.72	25.52	19.44	9.15	7.44	41.04	34.34	43.29	42.73	2.4	2.4	16,977
		A-Mem	**27.02**	**20.09**	**45.85**	**36.67**	**12.14**	**12.00**	**44.65**	**37.06**	50.03	49.47	**1.2**	**1.2**	2,520
	4o	LoCoMo	28.00	18.47	9.09	5.78	16.47	14.80	**61.56**	**54.19**	**52.61**	**51.13**	2.0	2.0	16,910
		ReadAgent	14.61	9.95	4.16	3.19	8.84	8.37	12.46	10.29	6.81	6.13	4.0	4.0	805
		MemoryBank	6.49	4.69	2.47	2.43	6.43	5.30	8.28	7.10	4.42	3.67	5.0	5.0	569
		MemGPT	30.36	22.83	17.29	13.18	12.24	11.87	60.16	53.35	34.96	34.25	2.4	2.4	16,987
		A-Mem	**32.86**	**23.76**	**39.41**	**31.23**	**17.10**	**15.84**	48.43	42.97	36.35	35.53	**1.6**	**1.6**	1,216
Qwen2.5	1.5b	LoCoMo	9.05	6.55	4.25	4.04	9.91	8.50	11.15	8.67	40.38	40.23	3.4	3.4	16,910
		ReadAgent	6.61	4.93	2.55	2.51	5.31	12.24	10.13	7.54	5.42	27.32	4.6	4.6	752
		MemoryBank	11.14	8.25	4.46	2.87	8.05	6.21	13.42	11.01	36.76	34.00	2.6	2.6	284
		MemGPT	10.44	7.61	4.21	3.89	13.42	11.64	9.56	7.34	31.51	28.90	3.4	3.4	16,953
		A-Mem	**18.23**	**11.94**	**24.32**	**19.74**	**16.48**	**14.31**	**23.63**	**19.23**	**46.00**	**43.26**	**1.0**	**1.0**	1,300
	3b	LoCoMo	4.61	4.29	3.11	2.71	4.55	5.97	7.03	5.69	16.95	14.81	3.2	3.2	16,910
		ReadAgent	2.47	1.78	3.01	3.01	5.57	5.22	3.25	2.51	15.78	14.01	4.2	4.2	776
		MemoryBank	3.60	3.39	1.72	1.97	6.63	6.58	4.11	3.32	13.07	10.30	4.2	4.2	298
		MemGPT	5.07	4.31	2.94	2.95	7.04	7.10	7.26	5.52	14.47	12.39	2.4	2.4	16,961
		A-Mem	**12.57**	**9.01**	**27.59**	**25.07**	**7.12**	**7.28**	**17.23**	**13.12**	**27.91**	**25.15**	**1.0**	**1.0**	1,137
Llama 3.2	1b	LoCoMo	11.25	9.18	7.38	6.82	11.90	10.38	12.86	10.50	51.89	48.27	3.4	3.4	16,910
		ReadAgent	5.96	5.12	1.93	2.30	12.46	11.17	7.75	6.03	44.64	40.15	4.6	4.6	665
		MemoryBank	13.18	10.03	7.61	6.27	15.78	12.94	17.30	14.03	52.61	47.53	2.0	2.0	274
		MemGPT	9.19	6.96	4.02	4.79	11.14	8.24	10.16	7.68	49.75	45.11	4.0	4.0	16,950
		A-Mem	**19.06**	**11.71**	**17.80**	**10.28**	**17.55**	**14.67**	**28.51**	**24.13**	**58.81**	**54.28**	**1.0**	**1.0**	1,376
	3b	LoCoMo	6.88	5.77	4.37	4.40	10.65	9.29	8.37	6.93	30.25	28.46	2.8	2.8	16,910
		ReadAgent	2.47	1.78	3.01	3.01	5.57	5.22	3.25	2.51	15.78	14.01	4.2	4.2	461
		MemoryBank	6.19	4.47	3.49	3.13	4.07	4.57	7.61	6.03	18.65	17.05	3.2	3.2	263
		MemGPT	5.32	3.99	2.68	2.72	5.64	5.54	4.32	3.51	21.45	19.37	3.8	3.8	16,956
		A-Mem	**17.44**	**11.74**	**26.38**	**19.50**	**12.53**	**11.83**	**28.14**	**23.87**	**42.04**	**40.60**	**1.0**	**1.0**	1,126

[1] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2025): 17577-17604.

A-Mem

- **Ablation:** removing Link Generation or Memory Evolution hurts performance
  - Both modules are necessary; they are complementary
- **Scaling:** retrieval time stays in microseconds even at 1 million memories
  - Same storage cost as baselines, but richer functionality

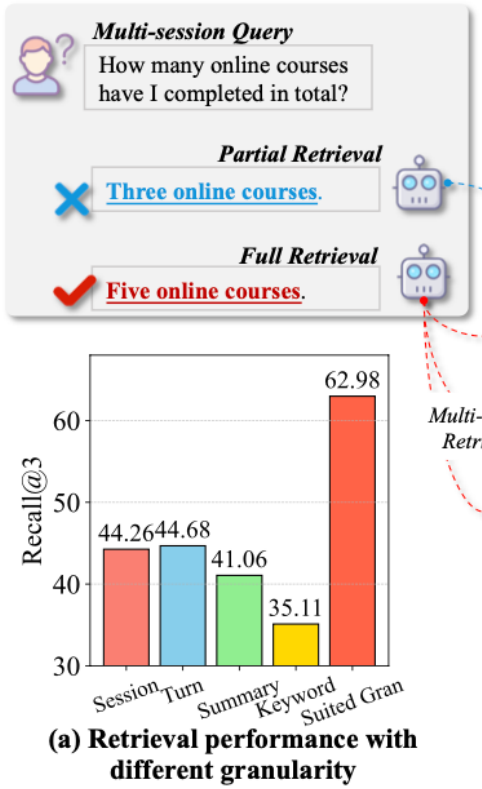
Method	Category									
	Multi Hop		Temporal		Open Domain		Single Hop		Adversial	
	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1
w/o LG & ME	9.65	7.09	24.55	19.48	7.77	6.70	13.28	10.30	15.32	18.02
w/o ME	21.35	15.13	31.24	27.31	10.13	10.85	39.17	34.70	44.16	45.33
A-MEM	**27.02**	**20.09**	**45.85**	**36.67**	**12.14**	**12.00**	**44.65**	**37.06**	**50.03**	**49.47**

Memory Size	Method	Memory Usage (MB)	Retrieval Time (μs)
1,000	A-MEM	1.46	0.31 ± 0.30
	MemoryBank [39]	1.46	0.24 ± 0.20
	ReadAgent [17]	1.46	43.62 ± 8.47
10,000	A-MEM	14.65	0.38 ± 0.25
	MemoryBank [39]	14.65	0.26 ± 0.13
	ReadAgent [17]	14.65	484.45 ± 93.86
100,000	A-MEM	146.48	1.40 ± 0.49
	MemoryBank [39]	146.48	0.78 ± 0.26
	ReadAgent [17]	146.48	6,682.22 ± 111.63
1,000,000	A-MEM	1464.84	3.70 ± 0.74
	MemoryBank [39]	1464.84	1.91 ± 0.31
	ReadAgent [17]	1464.84	120,069.68 ± 1,673.39

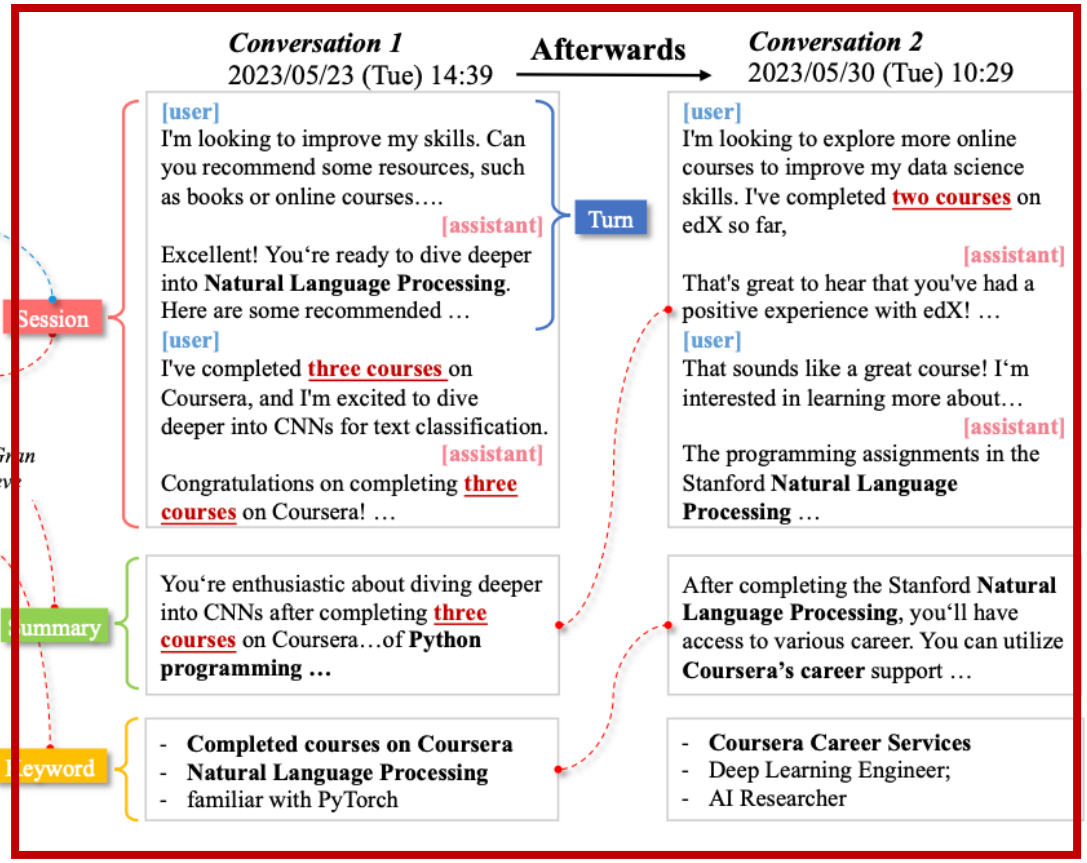
[1] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2025): 17577-17604.

MemGAS (Ours, ICLR 2026)

- **Problem**
  - LLM agents struggle with long conversations due to limited context
  - Existing memory systems use **single granularity** (session / turn / summary)
  - Partial retrieval or too much noise
- **Two Key Limits**
  - Insufficient cross-granularity memory connection



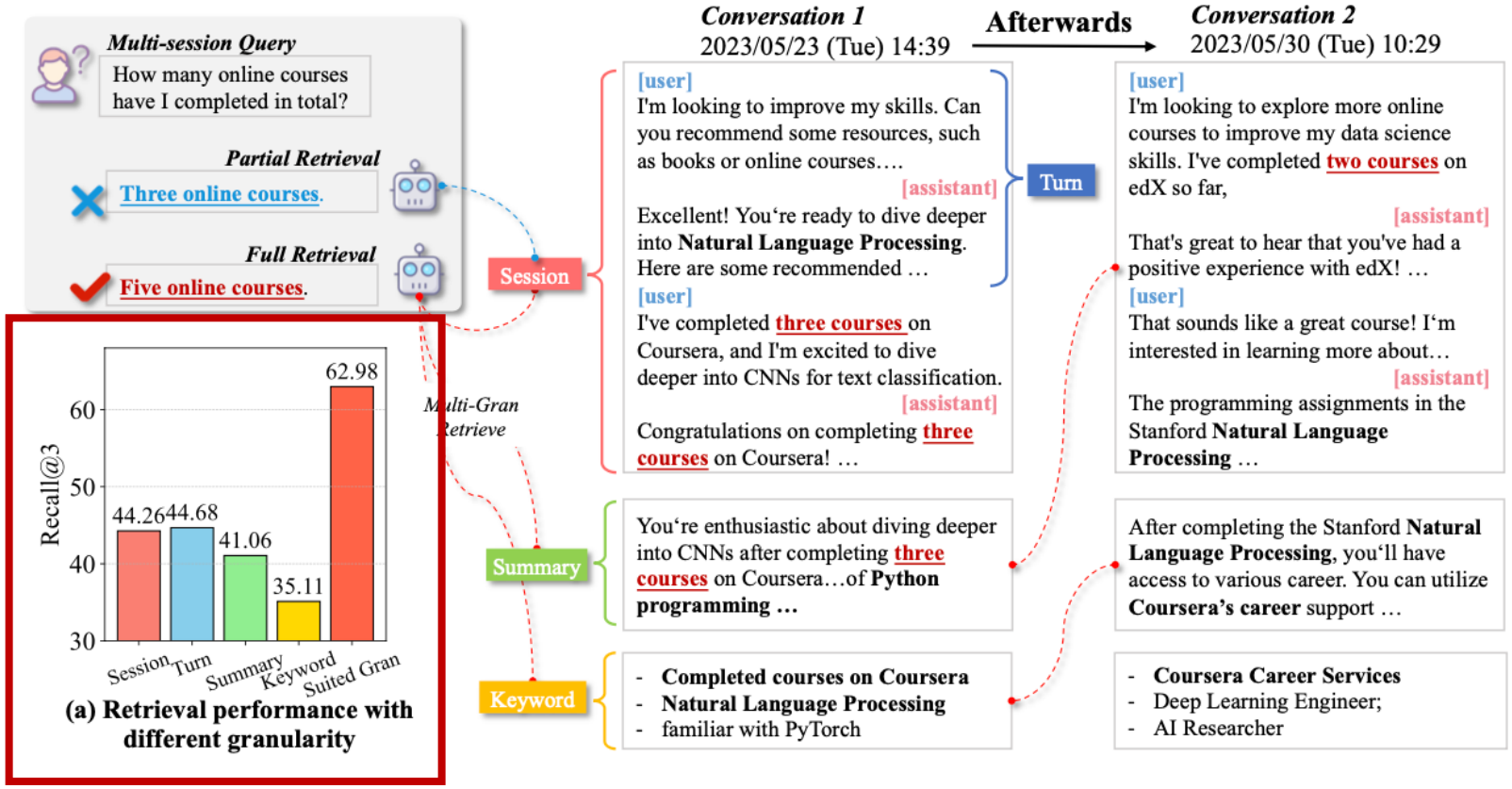
(a) Retrieval performance with different granularity



[1] Xu, Derong, et al. "From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents." ICLR (2026).

MemGAS (Ours, ICLR 2026)

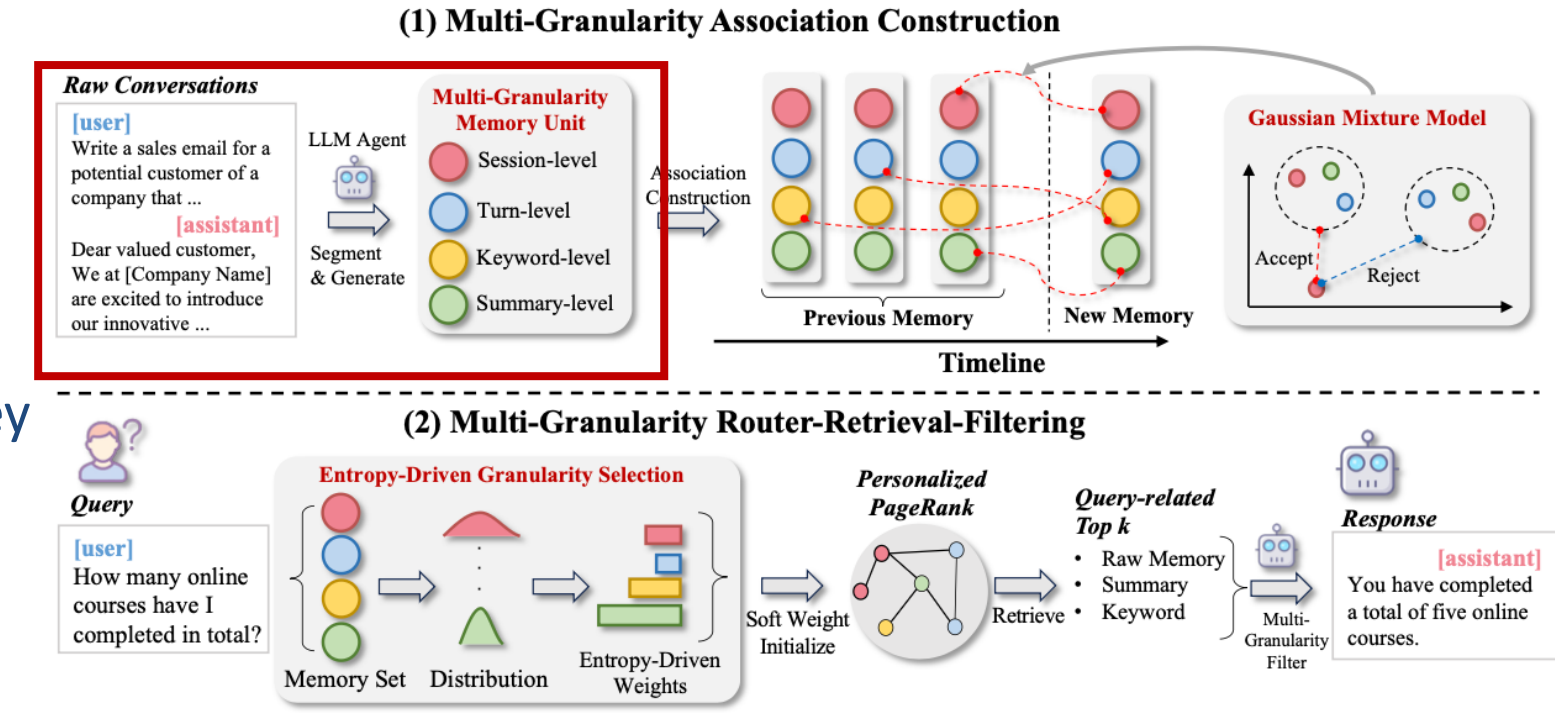
- **Problem**
  - LLM agents struggle with long conversations due to limited context
  - Existing memory systems use **single granularity** (session / turn / summary)
  - Partial retrieval or too much noise
- **Two Key Limits**
  - Insufficient cross-granularity memory connection
  - No adaptive selection of granularity for each query



[1] Xu, Derong, et al. "From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents." ICLR (2026).

MemGAS

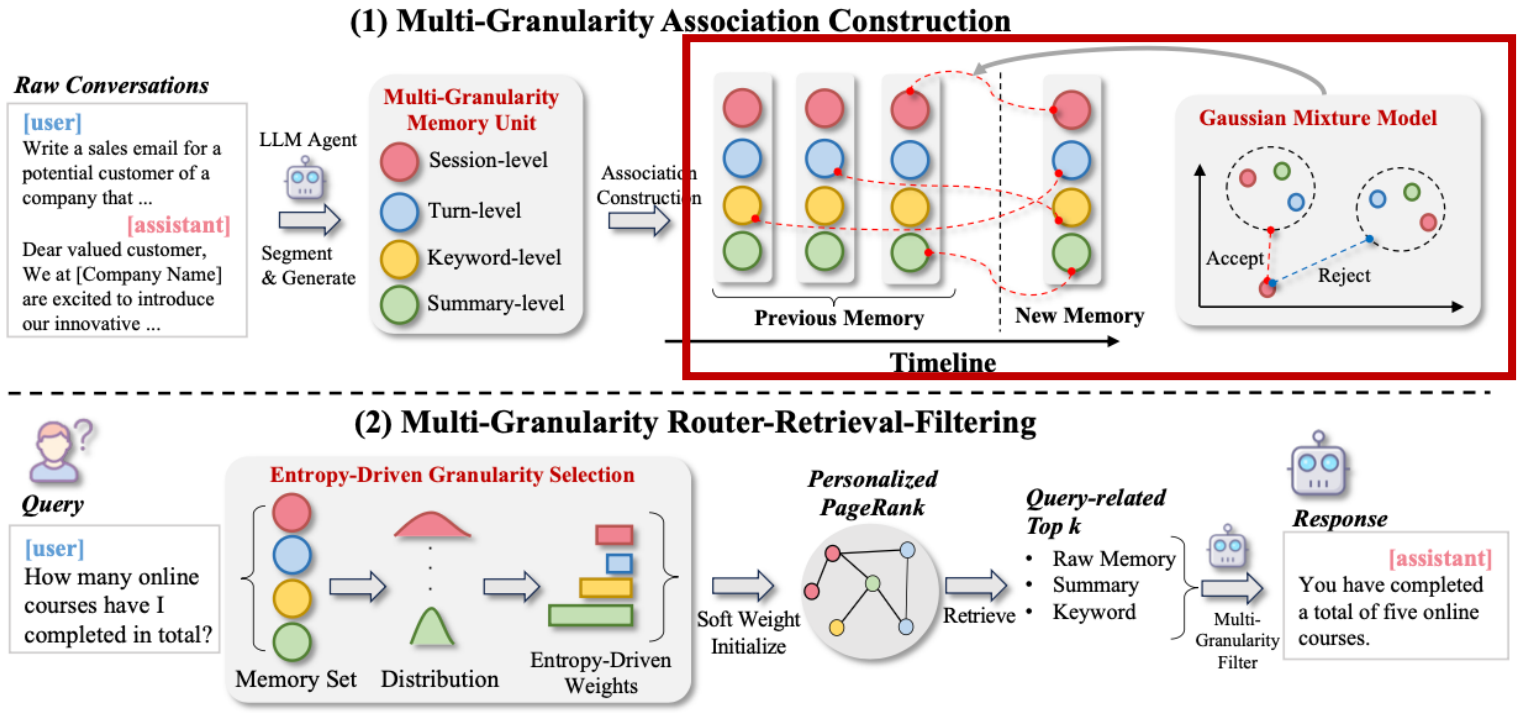
- **Step 1: Build Multi-Granularity Memory Units**
- For each session S_i, generate four levels:
  - **Session-level**: raw conversation
  - **Turn-level**: segmented dialogue turns
  - **Summary-level**: LLM-generated short summary
  - **Keyword-level**: LLM-extracted key concepts
- **Memory chunk:** M_i = {S_i, T_i, U_i, K_i}
- Each level is encoded as a **dense vector** for later retrieval



[1] Xu, Derong, et al. "From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents." ICLR (2026).

MemGAS

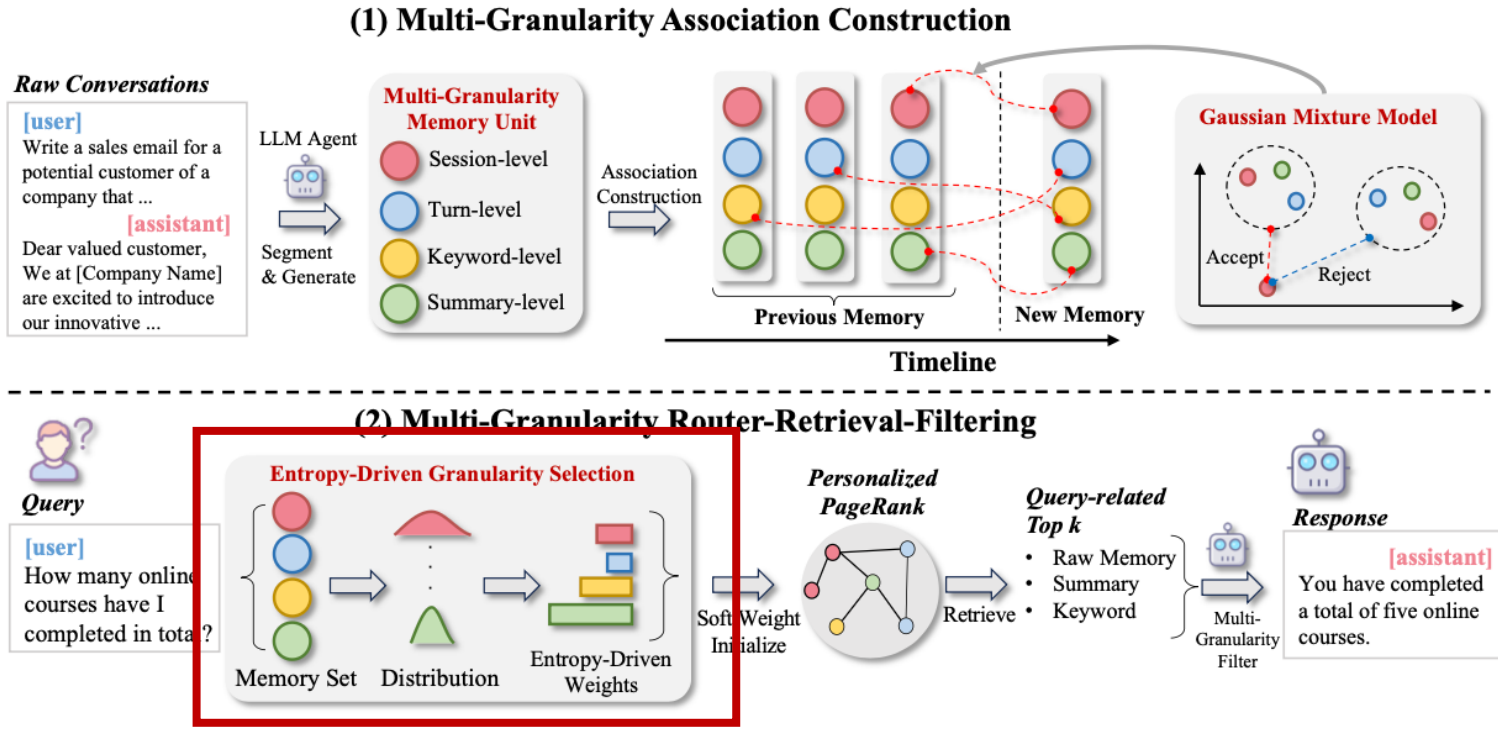
- **Step 2: Associate New Memory with Old Memories**
  - When a new memory M_new arrives: Compute similarity with all existing memories at every granularity
  - Cluster results into two sets using **GMM**
- Each granularity is a **node**; edges added between connected ones
  - One memory can link across different granularity levels
  - Builds an **association graph** that grows over time
- Mimics human-like memory consolidation



[1] Xu, Derong, et al. "From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents." ICLR (2026).

MemGAS

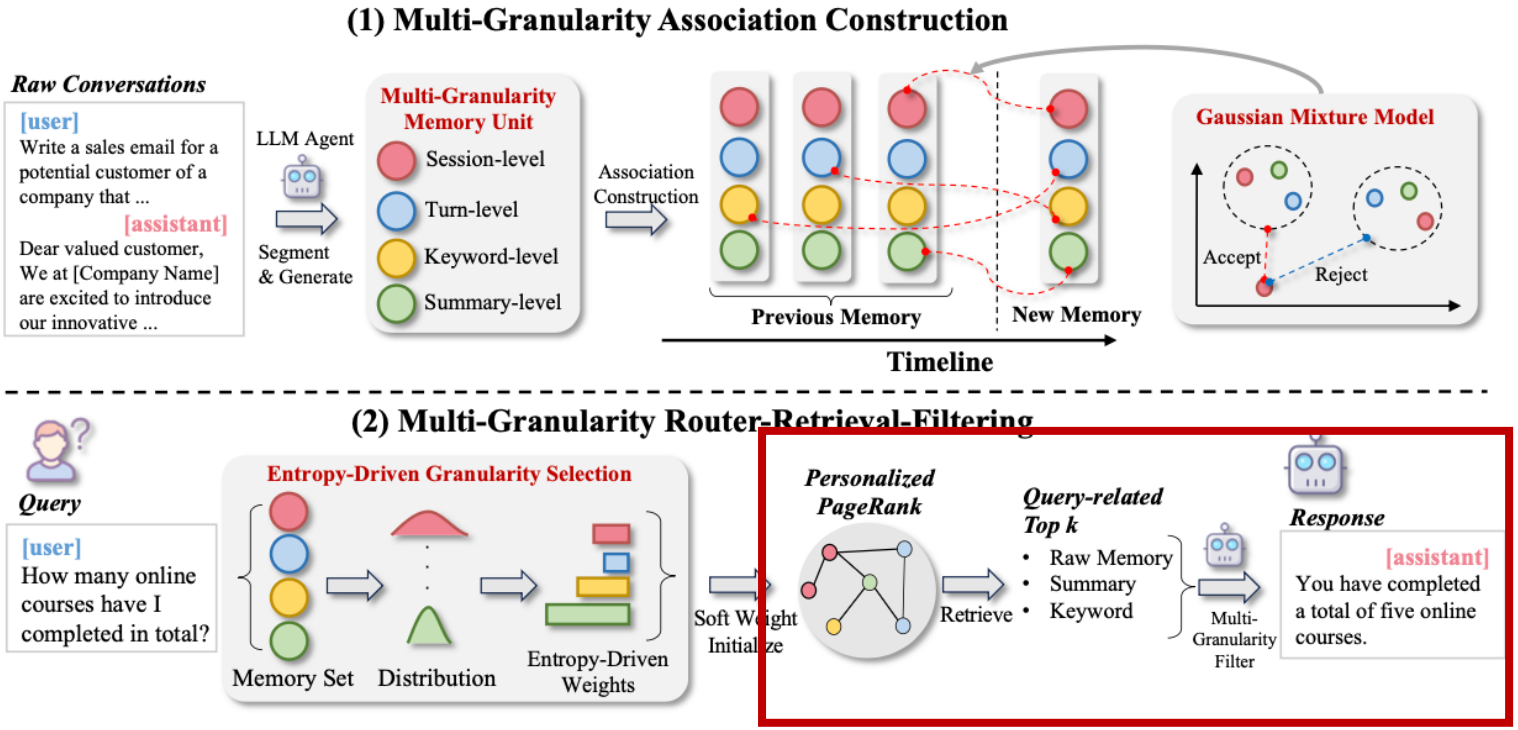
- **Step 3: Entropy-Driven Granularity Selection**
  - For each granularity g, compute similarity distribution + **Shannon entropy** H_g
  - **Lower entropy → higher confidence → larger weight**



[1] Xu, Derong, et al. "From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents." ICLR (2026).

MemGAS

- **Step 3: Entropy-Driven Granularity Selection**
  - For each granularity g, compute similarity distribution + **Shannon entropy** H_g
  - **Lower entropy → higher confidence → larger weight**
- **Step 4: Retrieval with Personalized PageRank**
  - Run **PPR** on the association graph → propagate relevance through links
  - Pick **top-K** nodes as candidate context
- **Step 5: LLM-based Redundancy Filter**
  - Remove irrelevant or repeated content before final answer



[1] Xu, Derong, et al. "From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents." ICLR (2026).

MemGAS

- **Setup**
  - **Datasets:** LoCoMo, Long-MT-Bench+, LongMemEval-s, LongMemEval-m
- **Findings**
  - MemGAS achieves **best results** on both QA and retrieval across most datasets
  - Especially strong on **multi-session** and **multi-hop** queries
  - Uses **competitive or fewer tokens** than structured baselines

Model	4o-J	F1	B-4	R-1	R-2	R-L	BS	Avg. Tokens	Avg. Latency
LongMemEval-s									
Full History	50.60	11.48	1.40	12.10	5.47	10.85	83.07	103,137	9.39
MPNet (Song et al., 2020)	53.20	13.96	2.21	14.49	6.78	12.93	83.72	8,173	1.82
Contriever (Izacard et al., 2021)	55.40	13.78	2.21	14.46	6.93	12.89	83.70	8,286	1.85
MPC (Lee et al., 2023)	53.80	13.60	1.74	14.27	6.49	12.95	83.49	8,457	2.66
RecurSum (Wang et al., 2025)	35.40	12.29	2.09	13.01	5.55	11.52	83.60	8,853	1.94
SeCom (Pan et al., 2025)	56.00	12.95	2.25	13.80	6.09	11.93	83.51	2,741	1.67
HippoRAG 2(Gutiérrez et al., 2025)	57.60	14.73	2.15	15.30	7.36	13.83	83.86	8,530	4.51
RAPTOR (Sarthi et al., 2024)	32.20	12.08	1.90	12.73	5.82	11.25	83.50	6,254	2.25
A-Mem (Xu et al., 2025b)	55.60	13.73	2.11	14.82	6.81	12.98	83.88	9,018	2.59
MemGAS (Ours)	**60.20**	**20.38**	**4.22**	**21.05**	**10.47**	**19.47**	**85.21**	8,829	2.55
LongMemEval-m									
Full History	12.20	5.70	0.78	6.27	2.08	5.28	81.62	128,000	12.88
MPNet (Song et al., 2020)	37.80	10.76	1.70	11.46	4.76	10.03	83.09	8,352	1.83
Contriever (Izacard et al., 2021)	42.80	11.88	1.66	12.56	5.63	11.02	83.28	8,467	1.92
MPC (Lee et al., 2023)	37.80	11.28	1.37	11.93	5.12	10.57	82.98	8,428	2.75
RecurSum (Wang et al., 2025)	23.80	10.04	1.70	10.89	4.26	9.21	83.12	8,927	1.99
SeCom (Pan et al., 2025)	42.80	11.33	1.79	12.03	5.07	10.49	83.36	2,821	1.63
MemGAS (Ours)	**45.40**	**16.85**	**3.39**	**17.60**	**8.25**	**16.14**	**84.69**	8,852	2.45
LoCoMo									
Full History	33.43	12.23	1.84	12.70	5.66	11.73	84.07	20,078	4.92
MPNet (Song et al., 2020)	38.07	14.44	2.35	14.90	6.83	13.90	84.42	2,472	1.29
Contriever (Izacard et al., 2021)	40.33	15.66	2.67	16.01	7.68	15.00	84.65	2,348	1.24
MPC (Lee et al., 2023)	40.38	14.81	1.99	15.10	6.83	14.13	84.43	2,683	1.95
RecurSum (Wang et al., 2025)	22.56	9.14	0.99	9.82	3.38	8.98	83.45	3,074	1.58
SeCom (Pan et al., 2025)	44.21	13.79	2.30	14.28	6.17	13.30	84.04	1,021	1.02
HippoRAG 2(Gutiérrez et al., 2025)	**45.62**	16.66	2.91	17.01	8.27	15.93	84.88	2,991	3.56
RAPTOR (Sarthi et al., 2024)	31.72	14.55	2.88	15.09	7.49	14.18	84.48	1,931	1.73
A-Mem (Xu et al., 2025b)	40.81	14.72	2.83	16.22	7.71	14.89	84.72	3,042	1.98
MemGAS (Ours)	41.07	**17.66**	**3.61**	**18.00**	**8.93**	**16.99**	**85.13**	2,825	1.88

Model	Recall@3	NDCG@3	Recall@5	NDCG@5	Recall@10	NDCG@10
LongMemEval-s						
MPNet (Song et al., 2020)	66.17	75.47	76.38	78.29	85.11	80.63
Contriever (Izacard et al., 2021)	71.06	79.72	81.28	82.47	90.00	84.29
RecurSum (Wang et al., 2025)	67.23	78.33	79.79	81.76	87.66	83.28
MPC (Lee et al., 2023)	60.00	70.90	68.09	73.27	80.00	76.59
SeCom (Pan et al., 2025)	71.06	80.88	80.43	83.08	89.15	85.11
HippoRAG 2(Gutiérrez et al., 2025)	75.53	85.44	84.68	87.32	91.28	88.73
MemGAS (Ours)	**78.51**	**86.83**	**88.94**	**88.77**	**94.47**	**89.96**
LongMemEval-m						
MPNet (Song et al., 2020)	37.02	48.68	45.74	52.51	61.28	56.55
Contriever (Izacard et al., 2021)	44.26	56.11	53.40	59.44	65.96	62.74
RecurSum (Wang et al., 2025)	23.19	32.10	31.70	36.99	43.62	41.40
MPC (Lee et al., 2023)	35.96	47.51	42.55	50.36	54.26	53.88
SeCom (Pan et al., 2025)	44.26	56.61	55.32	60.76	66.60	63.78
MemGAS (Ours)	**51.06**	**61.36**	**63.62**	**66.07**	**77.02**	**69.46**
LoCoMo						
MPNet (Song et al., 2020)	45.92	47.71	53.98	51.79	68.58	56.88
Contriever (Izacard et al., 2021)	49.90	52.15	58.26	56.29	71.80	60.92
RecurSum (Wang et al., 2025)	47.23	48.99	59.01	54.58	74.97	60.07
MPC (Lee et al., 2023)	49.50	51.47	57.45	55.53	71.85	60.47
SeCom (Pan et al., 2025)	55.24	57.90	64.80	62.36	78.30	66.97
HippoRAG 2(Gutiérrez et al., 2025)	56.00	58.57	65.00	62.50	78.05	66.7
MemGAS (Ours)	**57.30**	**58.76**	**67.32**	**63.62**	**81.82**	**68.4**

[1] Xu, Derong, et al. "From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents." ICLR (2026).

MemGAS

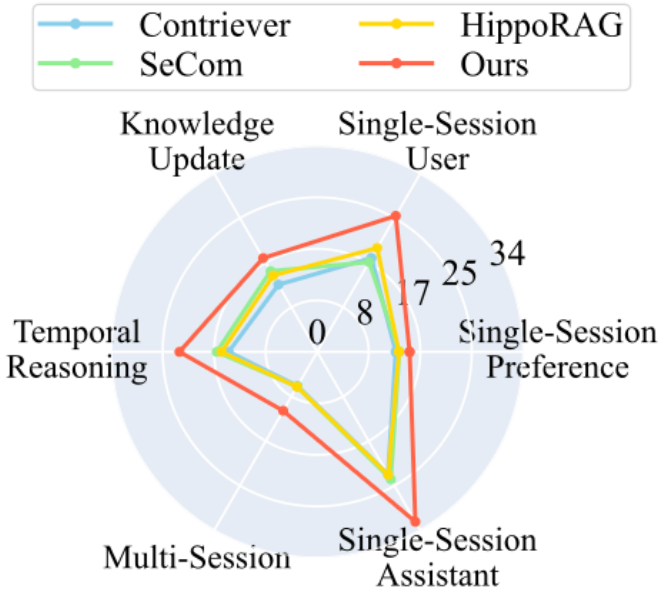
- **Ablation Study**
  - Removing **GMM, PPR, Memory Association, or Router** all hurt performance
  - All four modules are **necessary and complementary**
  - Extra latency from these modules is very small

Method	QA Performance					Retrieval Performance			
	GPT4o-J	F1	RogueL	Avg. Tokens	Total Latency (s)	R@3	R@5	R@10	Retrieval Latency (s)
MemGAS	**60.20**	**20.38**	**19.47**	8,829	2.5534	**78.51**	**88.94**	**94.47**	0.0239
w/o GMM	57.20	19.49	18.68	8,820	2.5506(Δ0.0028)	76.38	85.53	91.28	0.0232(Δ0.0007)
w/o PPR	56.60	19.76	18.85	8,772	2.5449(Δ0.0085)	75.96	85.96	90.64	0.0194(Δ0.0045)
w/o MA	56.80	17.69	19.00	8,734	2.5418(Δ0.0116)	74.89	85.74	91.49	0.0182(Δ0.0057)
w/o Router	56.60	18.88	18.62	8,763	2.5471(Δ0.0063)	75.53	85.74	92.34	0.0216(Δ0.0023)
w/o All	55.40	13.78	12.89	8,701	2.5343(Δ0.0191)	71.06	81.28	90.00	0.0160(Δ0.0079)

[1] Xu, Derong, et al. “From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents.” ICLR (2026).

MemGAS

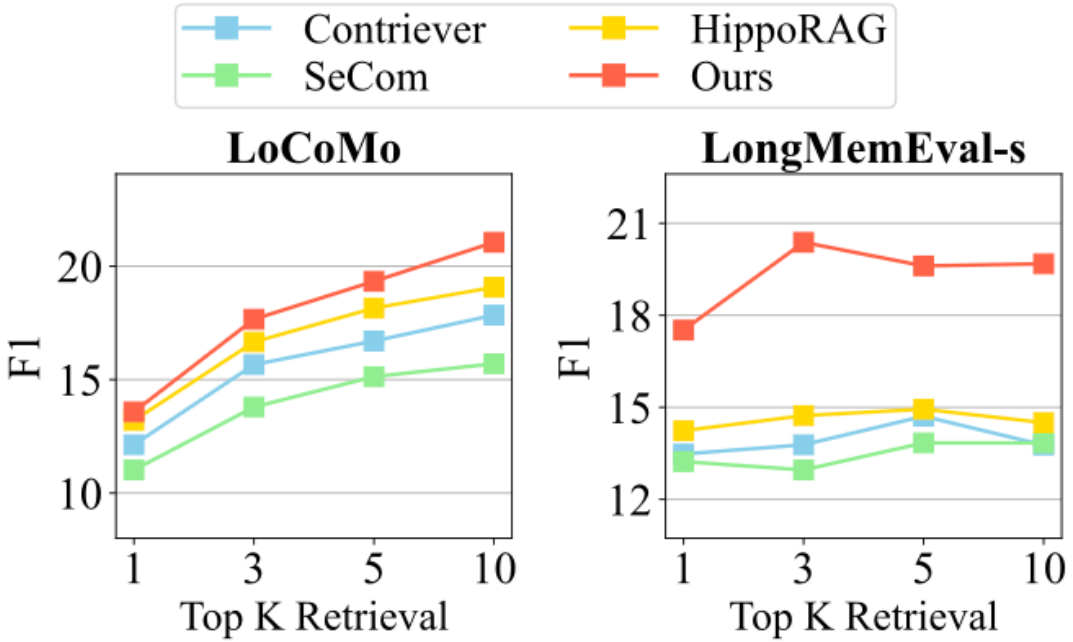
- **Ablation Study**
  - Removing **GMM, PPR, Memory Association, or Router** all hurt performance
  - All four modules are **necessary and complementary**
  - Extra latency from these modules is very small
- **Query Type Analysis**
  - Different queries prefer different granularities
  - The **router adapts** — close to the oracle "optimal selection"



[1] Xu, Derong, et al. “From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents.” ICLR (2026).

MemGAS

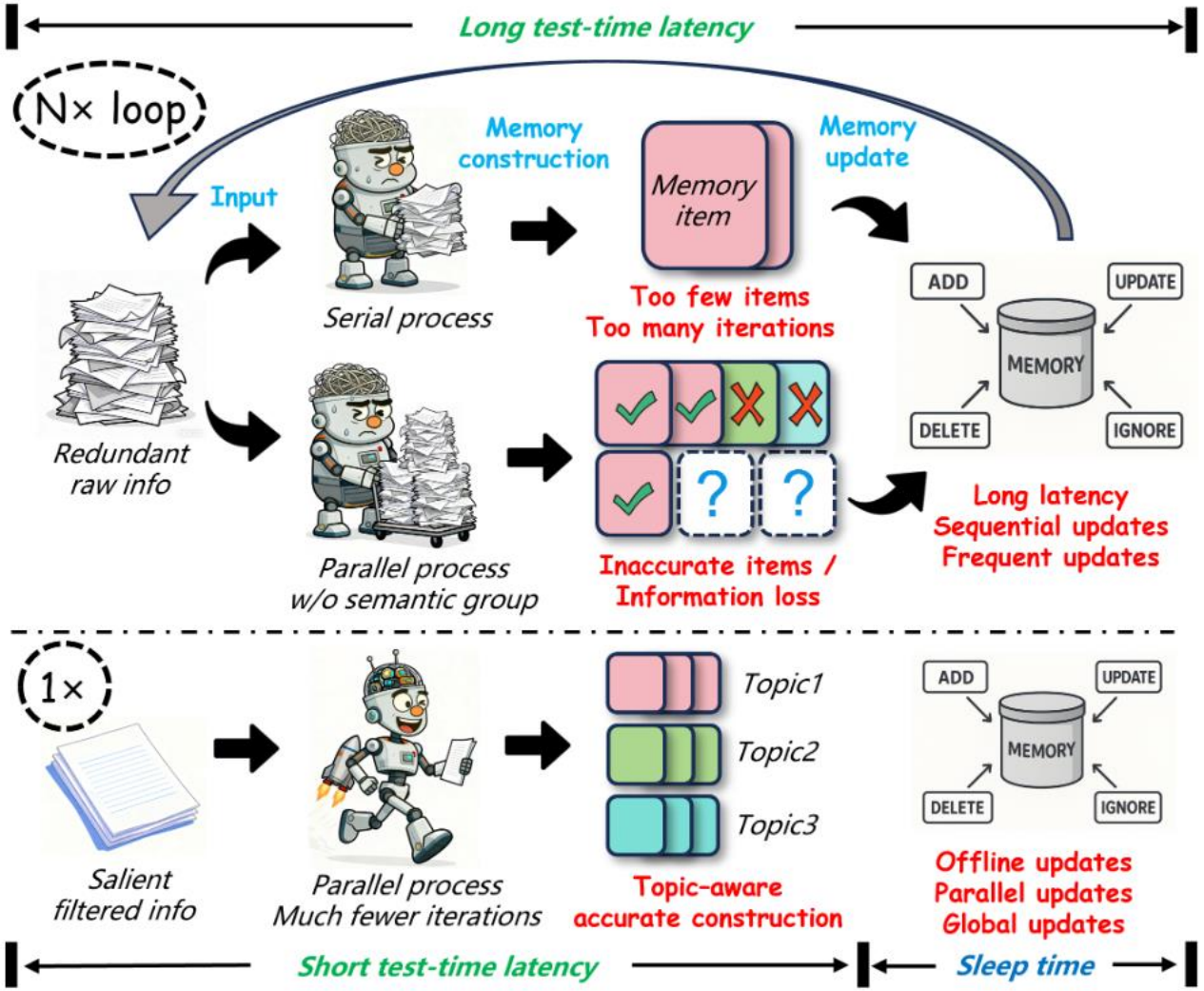
- **Ablation Study**
  - Removing **GMM, PPR, Memory Association, or Router** all hurt performance
  - All four modules are **necessary and complementary**
  - Extra latency from these modules is very small
- **Query Type Analysis**
  - Different queries prefer different granularities
  - The **router adapts** — close to the oracle "optimal selection"
- **Top-K Analysis**
  - MemGAS stays robust as K varies
  - Too large K introduces noise → filter helps



[1] Xu, Derong, et al. "From Single to Multi-Granularity: Toward Long-Term Memory Association and Selection of Conversational Agents." ICLR (2026).

LightMem (ICLR 2026)

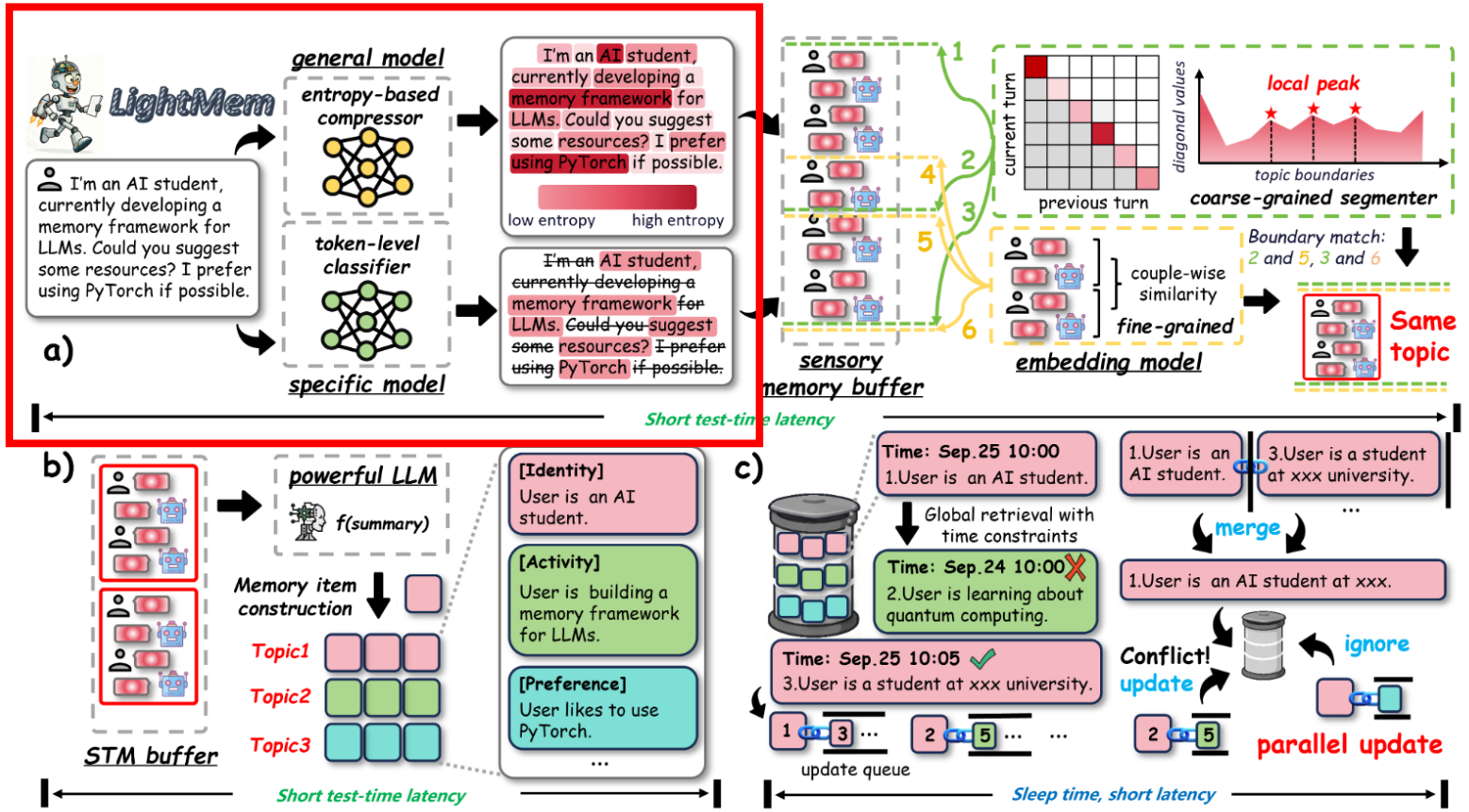
- **Background**
  - LLMs struggle with long, multi-turn interactions
  - Memory systems help LLMs keep persistent state
  - But existing memory systems are **expensive** and **slow**
- **Three Key Problems**
  - Raw input has lots of redundant tokens → wasted compute
  - Turn-by-turn processing → mixed topics, lost context
  - Online memory updates → high test-time latency



[1] Fang, Jizhan, et al. "Lightmem: Lightweight and efficient memory-augmented generation." arXiv preprint arXiv:2510.18866 (2025).ICLR (2026).

LightMem

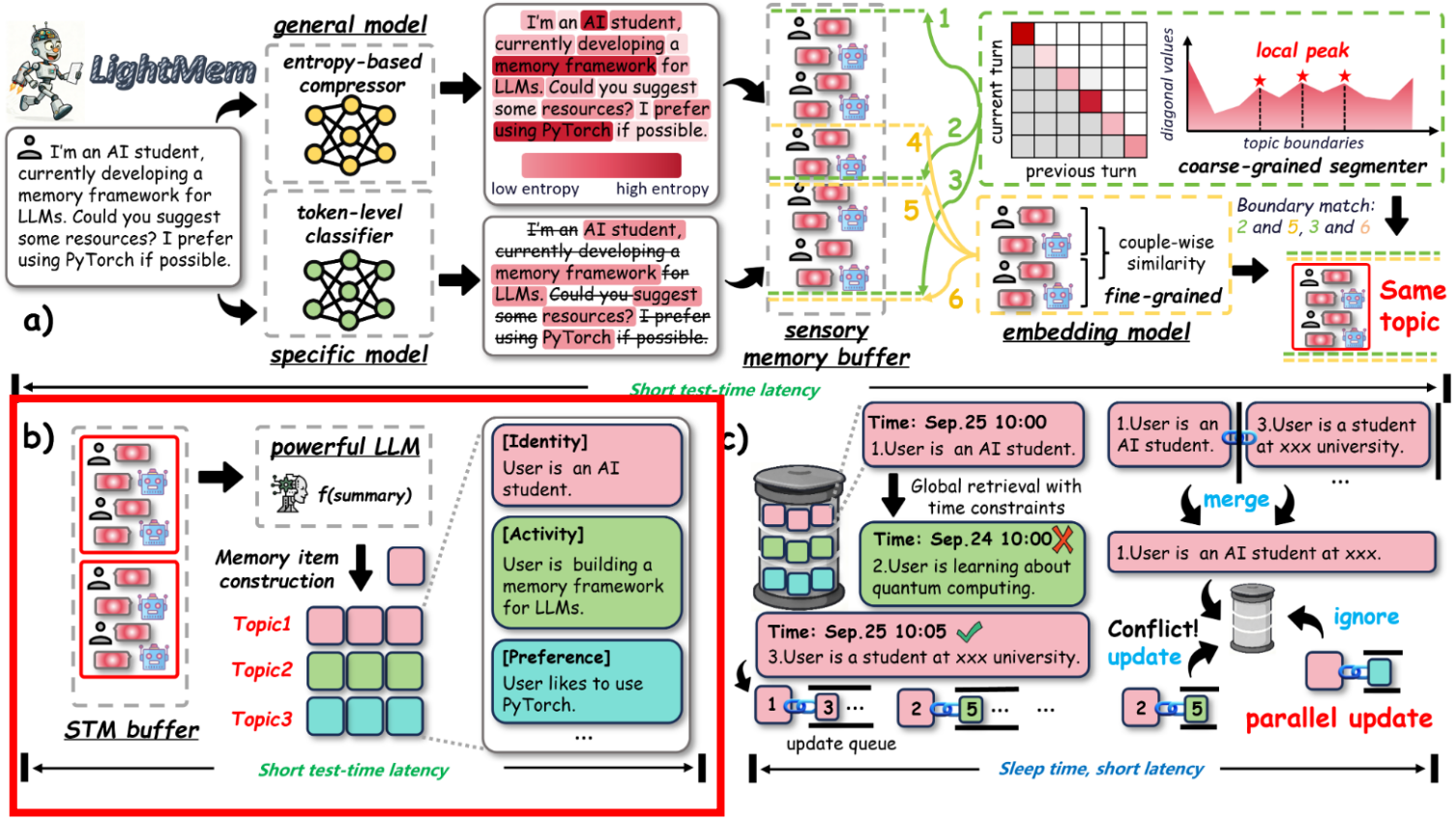
- **Sensory Memory: Filter the Noise First**
- **Step 1: Pre-Compression**
  - Use a small model (LLMLingua-2) to score each token
  - Keep only the high-value tokens
  - Drop redundant words → save tokens for later stages
- **Step 2: Topic Segmentation**
  - Store compressed input in a small buffer
  - When buffer is full, split into topics using:
    - Attention patterns (find local peaks)
    - Semantic similarity between turns
  - Final boundary = intersection of both signals



[1] Fang, Jizhan, et al. "Lightmem: Lightweight and efficient memory-augmented generation." arXiv preprint arXiv:2510.18866 (2025).ICLR (2026).

LightMem

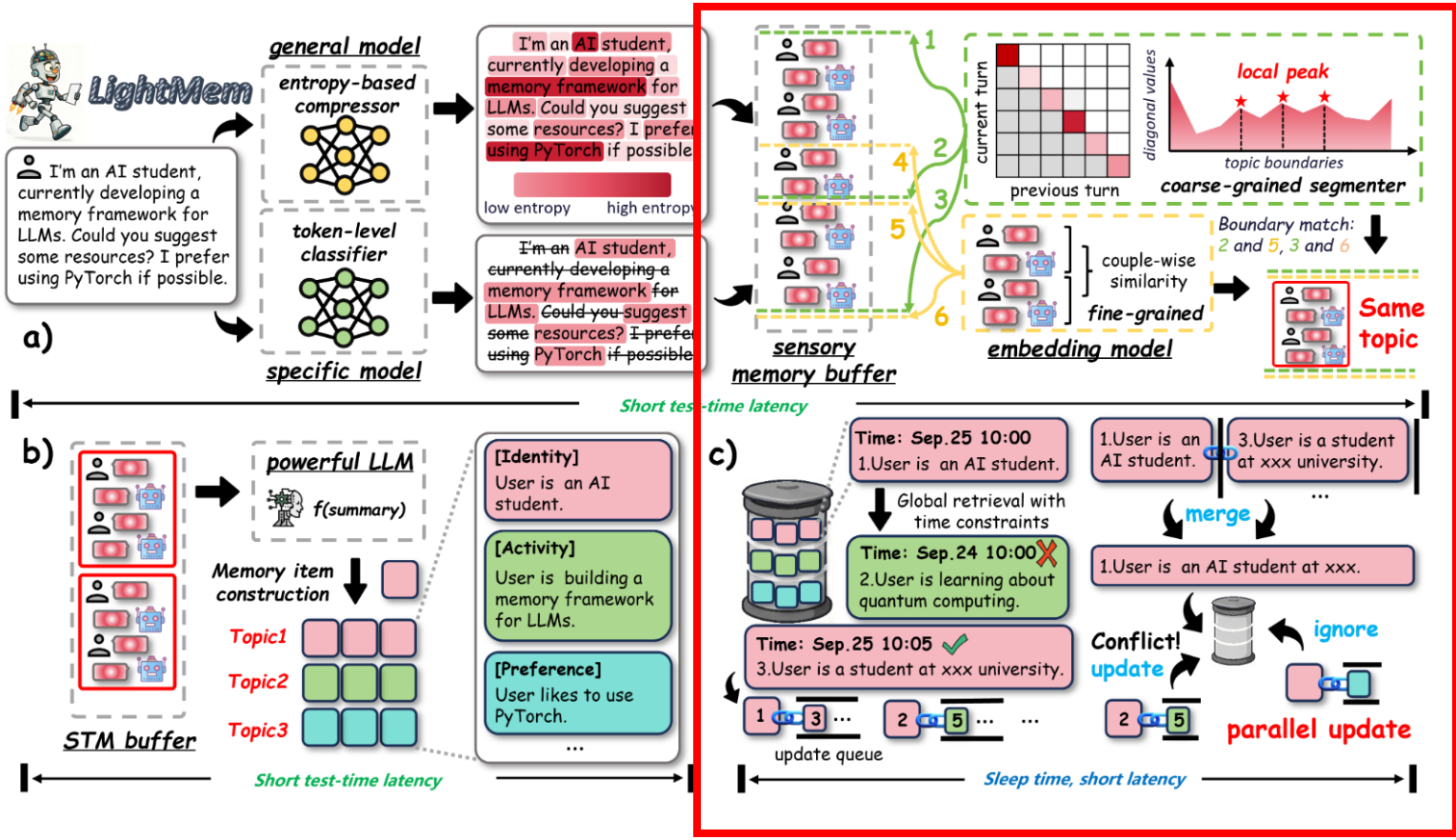
- **Topic-Aware Short-Term Memory**
- **The STM Buffer**
  - Receives topic-grouped segments from sensory memory
  - Holds them until token count hits a threshold (th)
- **Then: Summarize by Topic**
  - Call the LLM once per topic group
  - Generate a concise summary for each topic
  - Build a memory entry: {topic, summary, user, model}
- **Why this matters**
  - Turn-by-turn → too many LLM calls
  - Whole session → topics get mixed up
  - Topic-based → fewer calls + clean summaries



[1] Fang, Jizhan, et al. "Lightmem: Lightweight and efficient memory-augmented generation." arXiv preprint arXiv:2510.18866 (2025).ICLR (2026).

LightMem

- **Long-Term Memory with Sleep-Time Update**
- **At Test Time (Online): Soft Update**
  - New entries → just **insert** into LTM with a timestamp
  - No conflict checking, no LLM calls
  - Very fast, low latency for the user
- **At Sleep Time (Offline): Real Update**
  - For each entry, find top-k similar past entries
  - Only entries with later timestamps can update earlier ones
  - Build an **update queue** for each entry
  - **Run all queues in parallel**
- **Benefits**
  - No latency during conversation
  - Avoid wrong deletions from rushed LLM decisions
  - Parallel updates → much faster overall



[1] Fang, Jizhan, et al. "Lightmem: Lightweight and efficient memory-augmented generation." arXiv preprint arXiv:2510.18866 (2025).ICLR (2026).

LightMem

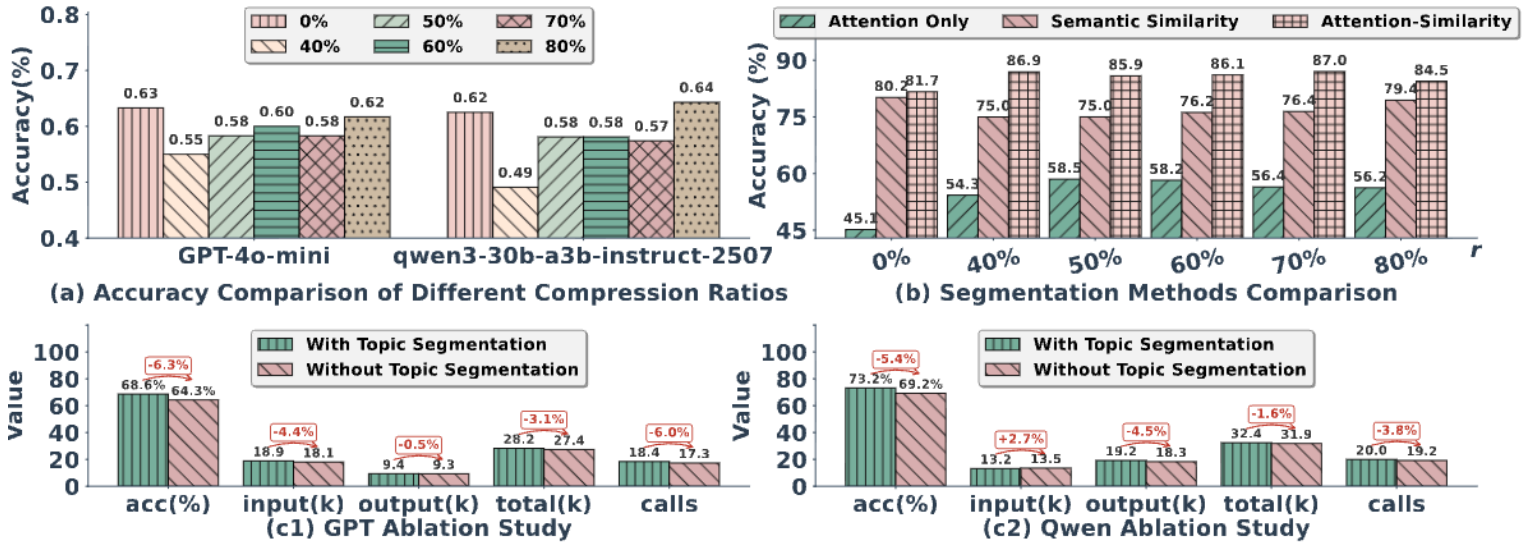
- **Effectiveness**
  - LightMem wins on accuracy across most settings
  - Works well on both GPT and Qwen backbones
- **Efficiency Gains**
  - Much lower token usage
  - Far fewer API calls
  - Faster runtime
  - Online-only cost is even lower than total cost

Method	ACC (%)	Summary Tokens (k)		Update Tokens (k)		Total (k)	Calls	Runtime (s)
		In	Out	In	Out			
GPT-4o-mini								
FullText	56.80	–	–	–	–	105.07	–	–
NaiveRAG	61.00	–	–	–	–	–	–	867.38
LangMem	37.20	–	–	982.68	119.48	1,102.16	520.62	2,293.70
A-MEM	62.60	214.66	42.82	1,157.52	190.81	1,605.81	986.55	5,132.06
MemoryOS	44.80	2,302.35	304.18	350.02	35.19	2,991.75	2,938.41	8,030.04
Mem0	53.61	424.13	17.76	560.17	150.56	1,152.62	811.57	4,248.49
LightMem								
r=0.5, th=256	64.29	20.80	10.01	–	–	30.81	25.67	302.69
(OP-update)	64.69	–	–	**44.46**	**2.56**	47.02	70.23	342.63
r=0.6, th=256	67.78	24.58	10.53	–	–	35.11	30.47	329.61
(OP-update)	65.39	–	–	53.98	3.18	57.16	85.07	411.56
r=0.7, th=512	**68.64**	**18.88**	**9.37**	–	–	**28.25**	**18.43**	**283.76**
(OP-update)	67.07	–	–	79.38	4.06	83.44	125.47	496.03
Qwen3-30B-A3B-Instruct-2507								
FullText	54.80	–	–	–	–	105.07	–	–
NaiveRAG	60.80	–	–	–	–	–	–	659.09
LangMem	50.80	–	–	1,311.96	118.06	1,430.02	495.12	3,237.16
A-MEM	65.20	219.21	66.98	1,260.54	318.20	1,864.93	989.30	5,367.51
MemoryOS	49.60	2,101.54	510.88	305.12	27.43	2,944.97	2,922.28	8,721.78
Mem0	39.51	424.20	**15.34**	411.50	111.35	1091.99	722.76	2,239.94
LightMem								
r=0.4, th=768	61.95	**9.01**	16.14	–	–	**25.15**	16.54	357.13
(OP-update)	62.34	–	–	111.13	7.88	119.01	176.02	1036.47
r=0.6, th=768	**70.20**	13.19	19.21	–	–	32.40	19.97	417.13
(OP-update)	65.14	–	–	**97.11**	**5.92**	103.03	152.93	1023.56
r=0.8, th=1024	68.69	14.82	18.49	–	–	33.31	**9.43**	**355.71**
(OP-update)	67.34	–	–	106.91	6.20	113.11	168.37	1026.90

[1] Fang, Jizhan, et al. "Lightmem: Lightweight and efficient memory-augmented generation." arXiv preprint arXiv:2510.18866 (2025).ICLR (2026).

LightMem

- **Pre-Compression Analysis**
  - Compression ratio 50%–80% → accuracy stays comparable
  - Tiny model, under 2GB GPU memory → negligible overhead
- **Topic Segmentation Ablation**
  - Remove it → accuracy drops noticeably
  - Hybrid (attention + similarity) beats single-signal methods



[1] Fang, Jizhan, et al. "Lightmem: Lightweight and efficient memory-augmented generation." arXiv preprint arXiv:2510.18866 (2025).ICLR (2026).

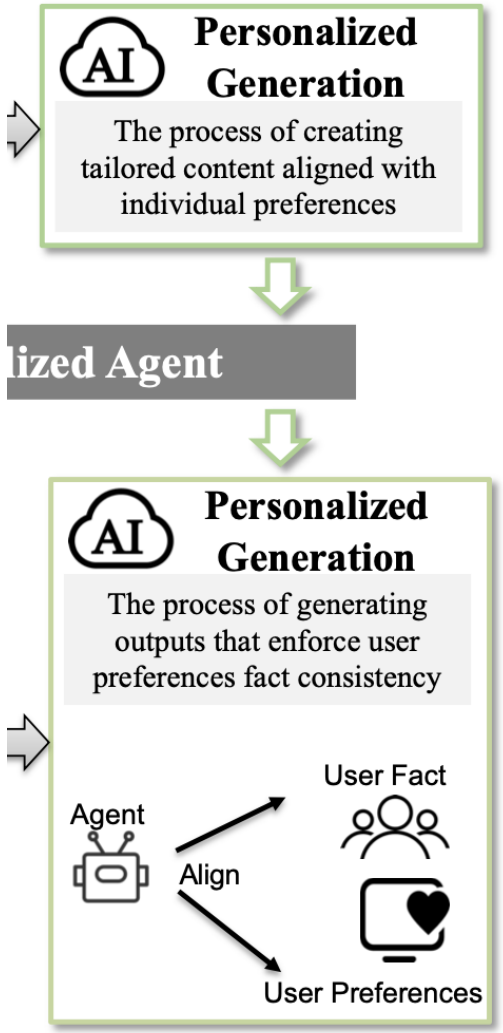
Personalized Generation

Motivation

- **Aligning outputs with user facts and preferences**, ensuring contextually relevant, trustworthy, and human-aligned responses.

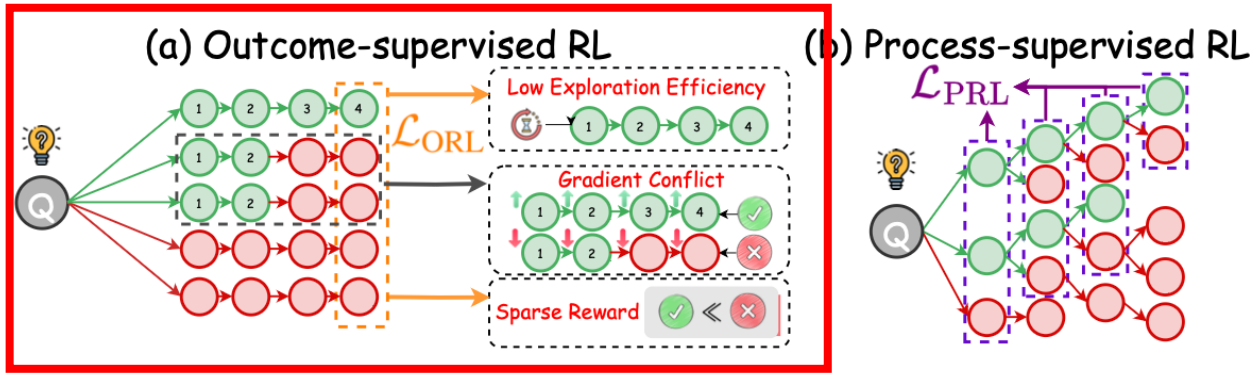
Category

- **Alignment with User Fact:**
  - Ensures factual accuracy and consistency in interactions, maintaining **trust** and **reliability** across various user.
- **Alignment with User Preferences:**
  - Adapts outputs to reflect **user values**, **personality traits**, and **interaction styles**, creating more personalized and engaging experiences.



ReasonRAG (Ours, NeurIPS 2025)

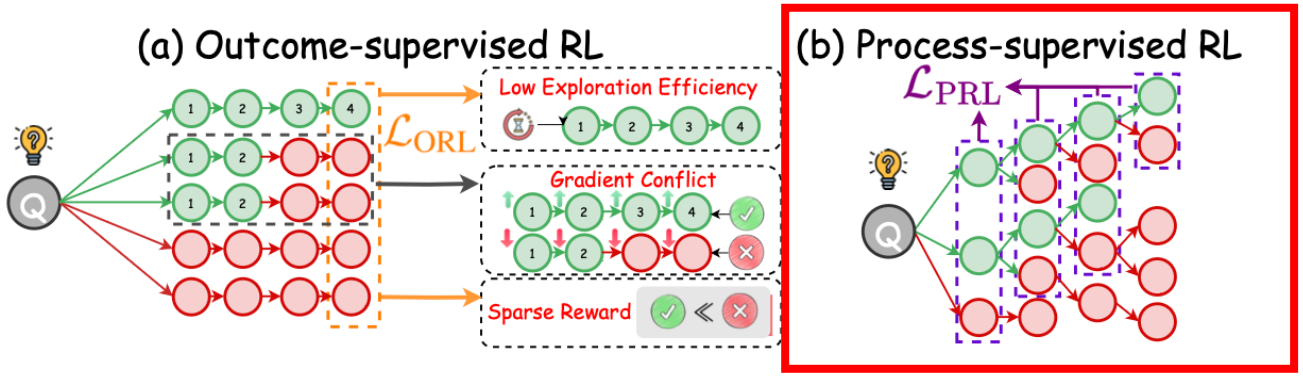
- **Background**
  - RAG helps LLMs use external knowledge
  - Agentic RAG goes further: dynamic retrieval, multi-step reasoning
  - Search-R1 uses **outcome-supervised RL** → final answer correctness as reward
- **Problems with Outcome-Supervised RL**
  - Low exploration efficiency — must finish whole chain to get reward
  - Gradient conflict — late errors penalize correct early steps
  - Sparse rewards — needs huge amounts of training data



[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

ReasonRAG (Ours, NeurIPS 2025)

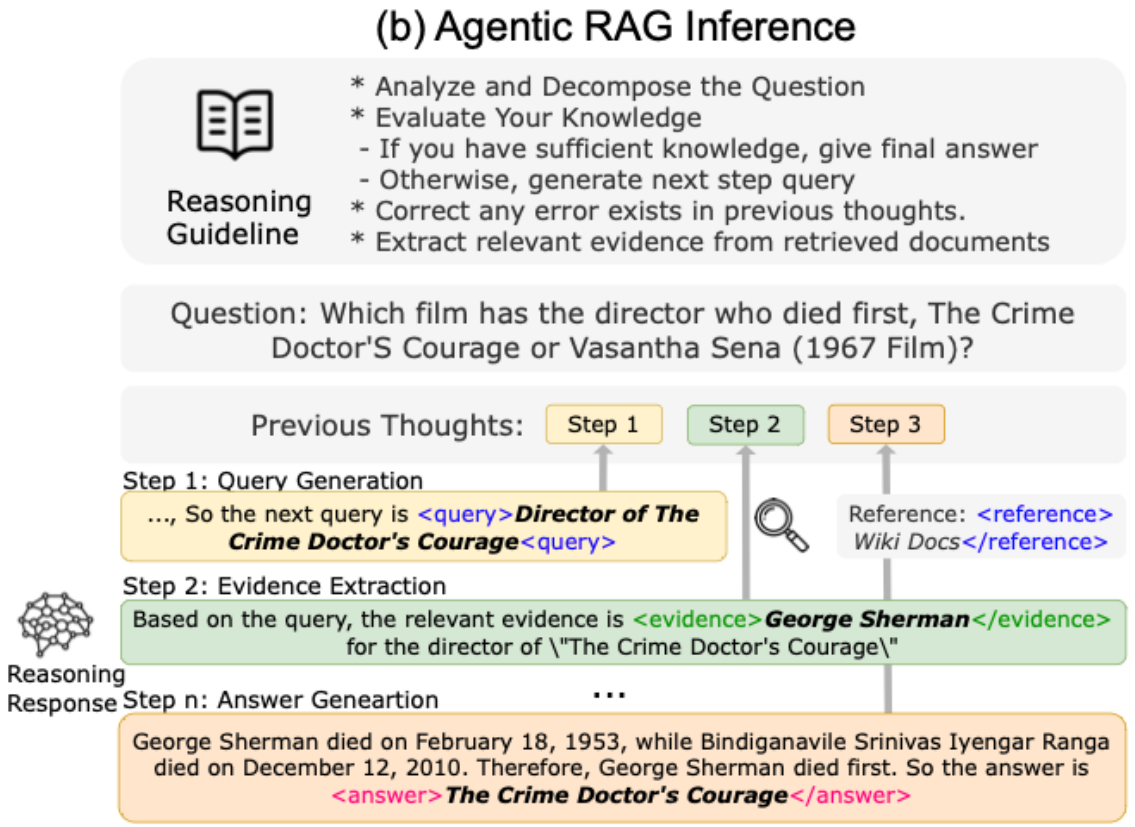
- **Background**
  - RAG helps LLMs use external knowledge
  - Agentic RAG goes further: dynamic retrieval, multi-step reasoning
  - Search-R1 uses **outcome-supervised RL** → final answer correctness as reward
- **Problems with Outcome-Supervised RL**
  - Low exploration efficiency — must finish whole chain to get reward
  - Gradient conflict — late errors penalize correct early steps
  - Sparse rewards — needs huge amounts of training data
- **Our Idea: Process-Supervised RL**
  - Give rewards at **each step**, not only at the end
  - Denser feedback → faster, more stable learning



[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

ReasonRAG

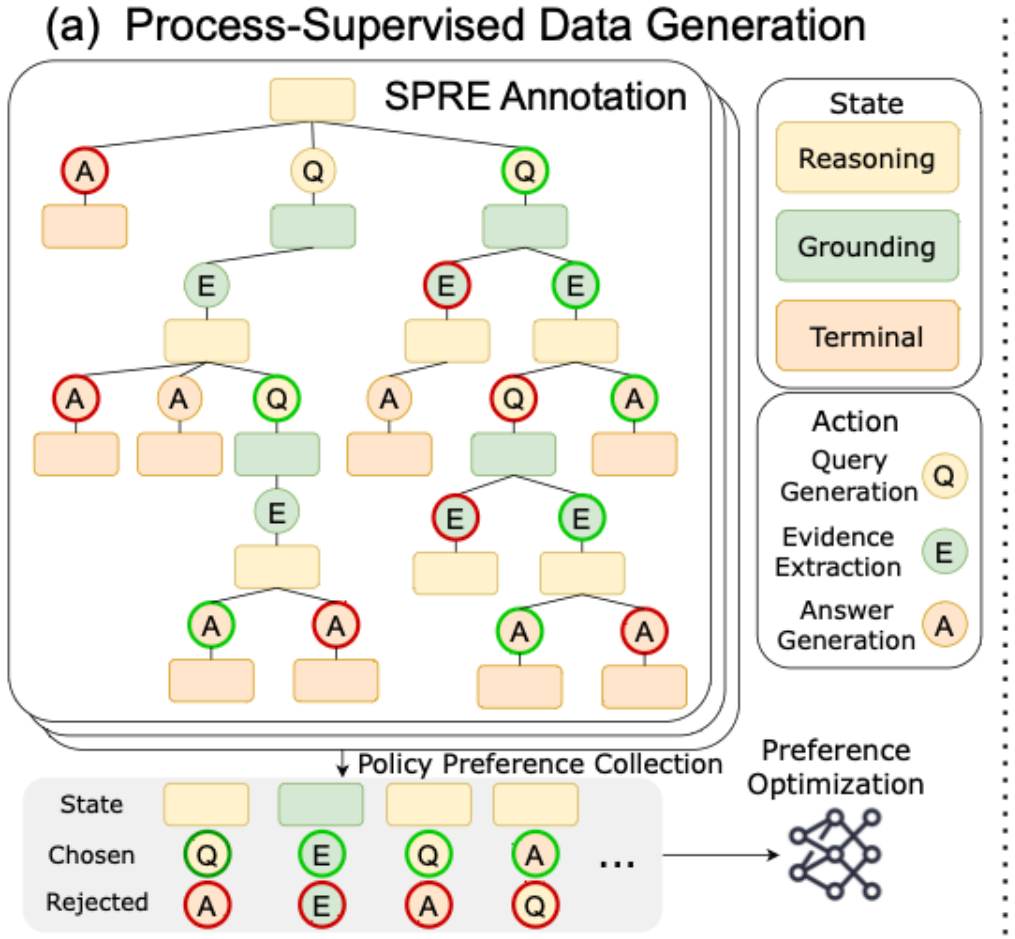
- **Agentic RAG Workflow** (3 states)
  - **Reasoning** → generate query OR final answer
  - **Grounding** → extract evidence from retrieved docs
  - **Terminal** → answer ready, stop
- **Three Core Actions**
  - Query Generation
  - Evidence Extraction
  - Answer Generation
- **SPRE: Shortest Path Reward Estimation**
  - Problem: no ground-truth reward for intermediate steps
  - Solution: simulate multiple continuations (rollouts) from each step
  - Score = correctness × **decay factor** for length
  - Rewards short, correct paths; penalizes long detours



[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

ReasonRAG

- **Our Solution: Monte Carlo Tree Search (MCTS)**
- **Three Core Operations**
  - **Selection** — pick promising nodes using UCB score (balance exploration & exploitation)
  - **Expansion** — sample a new action from the LLM policy
  - **Backpropagation** — update node rewards via SPRE
- **Tree Structure**
  - Root = original question
  - Each node = one state (question + previous thoughts + stage)
  - Each edge = one action (query / evidence / answer)



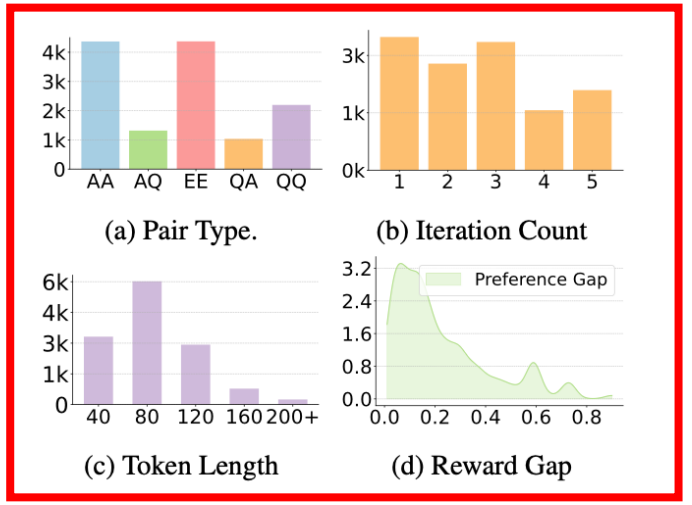
[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

ReasonRAG

- **Building RAG-ProGuide**
  - Source: PopQA, HotpotQA, 2WikiMultihopQA
  - Run MCTS on each question → collect tree-structured data
  - Post-processing: remove duplicates & low-gap pairs
- **Dataset Highlights**
  - Balanced over 3 actions: query / evidence / answer
  - Covers single-hop and multi-hop reasoning
  - Rich preference-pair diversity
- **Training: Process-Supervised DPO**
  - Take preference pairs (chosen vs. rejected) from the tree
  - Apply Direct Preference Optimization step by step
  - Model learns to prefer **short, correct** reasoning trajectories
- **Outcome:** LLM learns to autonomously
  - Decide when to search
  - Write good queries
  - Pick relevant evidence
  - Synthesize the final answer

Statistics	Number
Questions	4603
- PopQA	704 (15.3%)
- HotpotQA	2843 (61.8%)
- 2WikiMultihopQA	1056 (22.9%)
Actions	13289
- Query Generation	3295 (24.8%)
- Evidence Extraction	4305 (32.4%)
- Answer Generation	5689 (42.8%)
Avg./Min./Med./Max. Iteration	2.7/1/3/5
Avg./Min./Med./Max. Tokens	65.5/9/60/625

Table 1: Overall Dataset Statistics



[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

Experiments

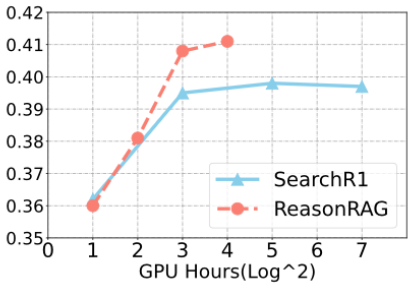
- **Main Findings**
  - **Data efficient** — only 5k queries, beats Search-R1's 90k
  - **Strong on multi-hop** — bigger gains on complex reasoning
  - **Good generalization** — wins on out-of-domain benchmarks too

Type	Method	PopQA		HotpotQA		2WikiMulti		Bamboogle		MuSiQue		Avg.	
		EM	F_1	EM	F_1	EM	F_1	EM	F_1	EM	F_1	EM	F_1
Zero-shot	Naïve Generation	12.7	16.5	15.7	24.8	20.2	28.0	6.4	17.4	2.7	10.2	11.5	19.4
	Standard RAG	38.4	44.7	29.3	39.9	29.4	36.3	17.6	24.1	6.7	15.1	24.3	32.0
Active	FLARE	14.3	17.6	18.1	25.7	27.9	32.8	12.0	20.8	4.3	12.6	15.3	21.9
	Self-RAG(146k)	22.7	33.9	21.0	29.7	12.0	25.2	1.6	10.9	4.6	13.3	12.4	22.6
Adaptive	AdaptiveRAG(3k)	36.6	41.5	29.1	40.7	24.2	33.4	18.4	26.1	6.9	14.3	23.0	31.2
RAG-CoT	Iter-Retgen	38.7	44.9	30.3	42.1	31.2	38.7	19.2	26.4	7.7	14.2	25.4	33.3
	IRCoT	36.2	43.6	27.7	41.5	23.5	32.5	17.2	22.5	8.6	13.2	22.6	30.7
Summary	RECOMP	40.5	45.8	29.7	41.2	33.2	39.4	21.7	28.6	9.2	15.8	26.9	34.2
	LongLLMLingua	39.2	45.1	31.4	43.2	34.5	40.2	20.3	27.4	8.7	14.9	26.8	34.2
	Selective-Context	34.9	41.5	19.3	27.3	20.3	29.7	15.3	22.6	6.1	13.7	19.2	27.0
Reasoning	Search-o1	33.2	40.3	24.8	38.1	16.4	27.1	30.4	40.6	6.3	13.7	22.2	31.96
	AutoRAG(10k)	38.6	44.1	33.3	43.7	39.5	46.1	24.8	32.2	11.3	18.3	29.5	36.9
	Search-R1(90k)	39.7	44.8	37.0	47.0	41.4	48.0	32.0	43.8	**14.6**	19.9	32.8	40.7
	ReasonRAG(5k)	**41.5***	**46.2***	**38.4***	**48.9***	**43.6***	**50.4***	**36.0***	**45.5***	12.8	**20.6***	**34.4***	**42.3***

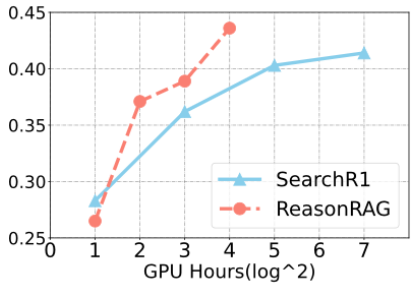
[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

Experiments

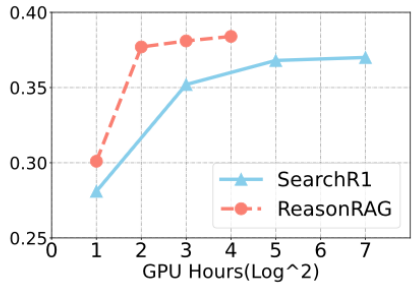
- **Training Efficiency**
  - ReasonRAG reaches strong performance with **far fewer GPU hours**
  - Gap is biggest on multi-hop tasks
  - Process rewards = faster convergence



(a) PopQA



(b) 2WikiMultiHopQA



(c) HotpotQA

[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

ReasonRAG

- **Process vs. Outcome vs. SFT**
  - **PRL** (ours) — best across all datasets
  - **ORL** (outcome) — second best, but needs **2× data** to catch up
  - **SFT** — overfits to multi-hop paths, hurts single-hop
  - **Base** — no training, weakest

Table 3: Impact of different optimization strategies on ReasonRAG 's effectiveness.

Method	PopQA		HotpotQA		2WikiMulti		Bamboogle		MuSiQue		Avg.	
	EM	F_1	EM	F_1	EM	F_1	EM	F_1	EM	F_1	EM	F_1
ReasonRAG (Base)	35.6	42.7	23.7	38.2	15.2	28.9	28.0	38.7	7.7	15.4	22.0	32.8
ReasonRAG (SFT)	31.6	37.4	26.8	38.7	35.1	40.9	17.6	27.3	8.6	15.5	23.9	32.0
ReasonRAG (RL-ORL): 5k queries	23.0	30.9	28.1	32.6	32.0	43.8	17.5	24.1	5.9	13.1	21.3	28.9
ReasonRAG (RL-ORL): 10k queries	39.5	45.7	36.7	46.7	40.5	47.2	30.7	40.6	12.6	19.5	32.0	39.9
ReasonRAG (RL-PRL)	**41.5**	**46.2**	**38.4**	**48.9**	**43.6**	**50.4**	**36.0**	**45.5**	**12.8**	**20.6**	**34.5**	**42.3**

[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

ReasonRAG

- **Retrieval Steps Analysis**
  - Performance grows with more retrieval steps, then saturates
  - PopQA → converges in 2–3 steps (simple)
  - HotpotQA / 2WikiMulti → needs 3–5 steps (complex)

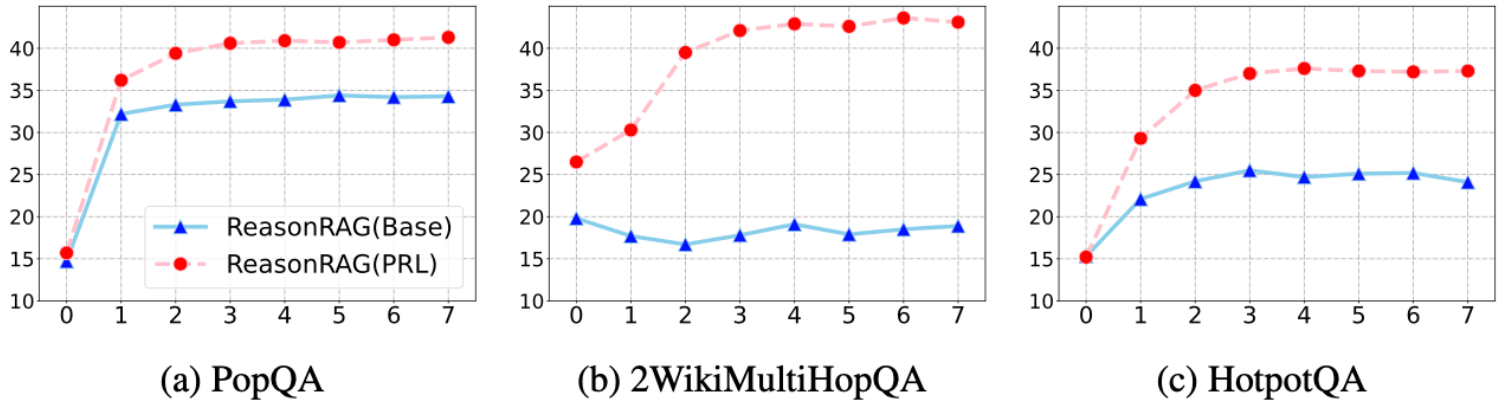
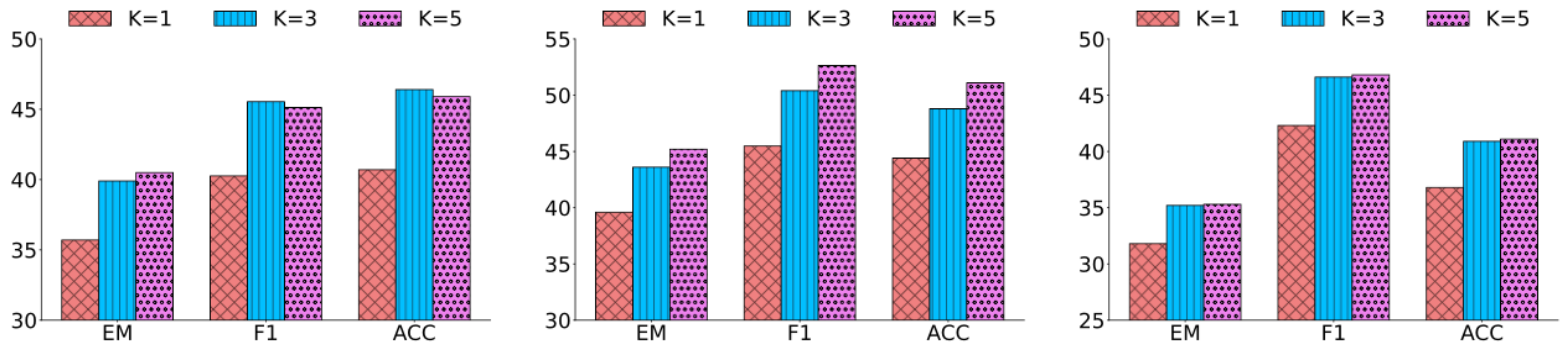


Figure 6: The EM performance across varying retrieval iterations on 3 benchmarks.

[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

ReasonRAG

- Model **adapts depth to question complexity**
- **Top-k Documents Analysis**
  - k=1 → too little context, hurts reasoning
  - k=3 → sweet spot for most datasets
  - k=5 → helpful for hard multi-hop questions



(a) PopQA (b) 2WikiMultiHopQA (c) HotpotQA

Figure 7: Effect of top-k retrieved documents on ReasonRAG's performance across 3 datasets.

[1] Zhang, Wenlin, et al. "Process vs. outcome reward: Which is better for agentic rag reinforcement learning." Advances in Neural Information Processing Systems 38 (2025): 58701-58729.

Agenda

1. Introduction
9:00 – 9:50



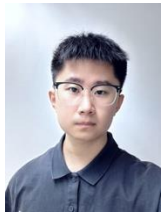
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

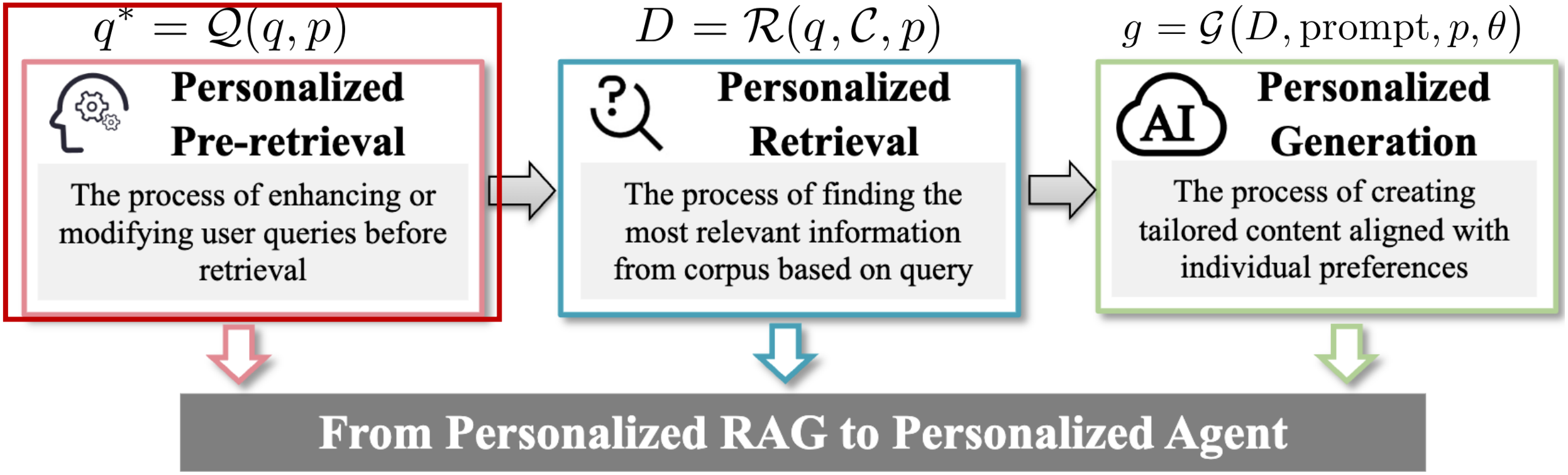
8. Hands on Projects
16:45 – 17:30



Pengyue Jia

Data & Evaluation Methodology

➢Data & Evaluation Methodology in different stages:

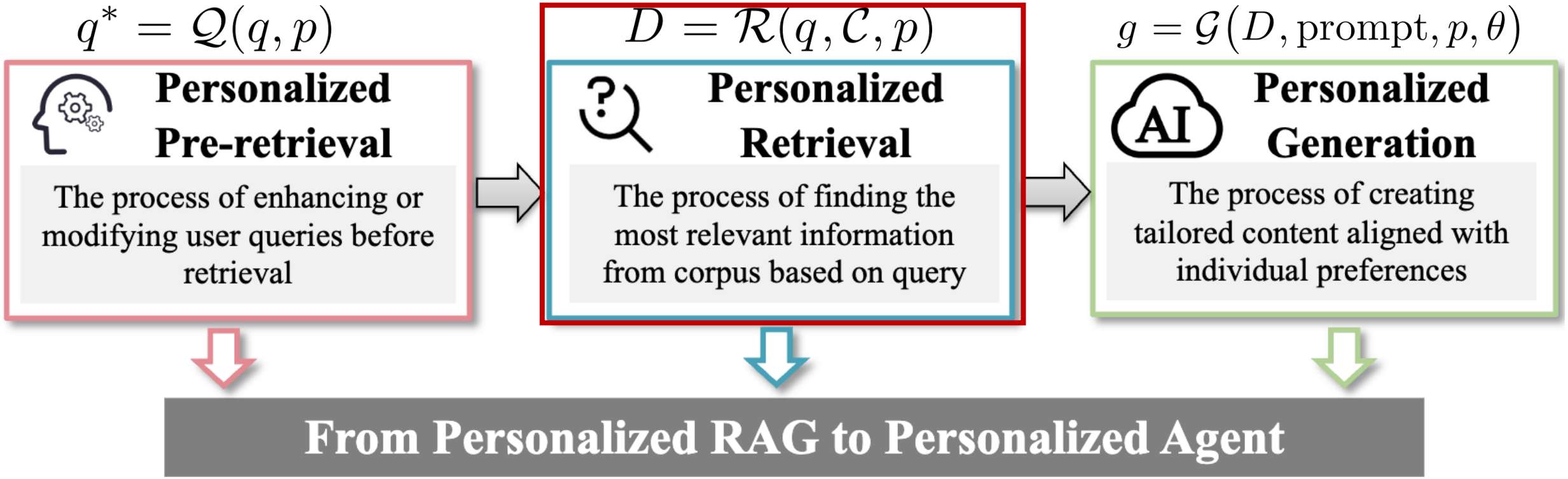


Proxy of the whole work flow.

$$g = \mathcal{G}\left(\mathcal{R}\left(\mathcal{Q}\left(q, p\right), \mathcal{C}, p\right), \text{prompt}, p, \theta\right)$$

Data & Evaluation Methodology

➢ Data & Evaluation Methodology in different stages:

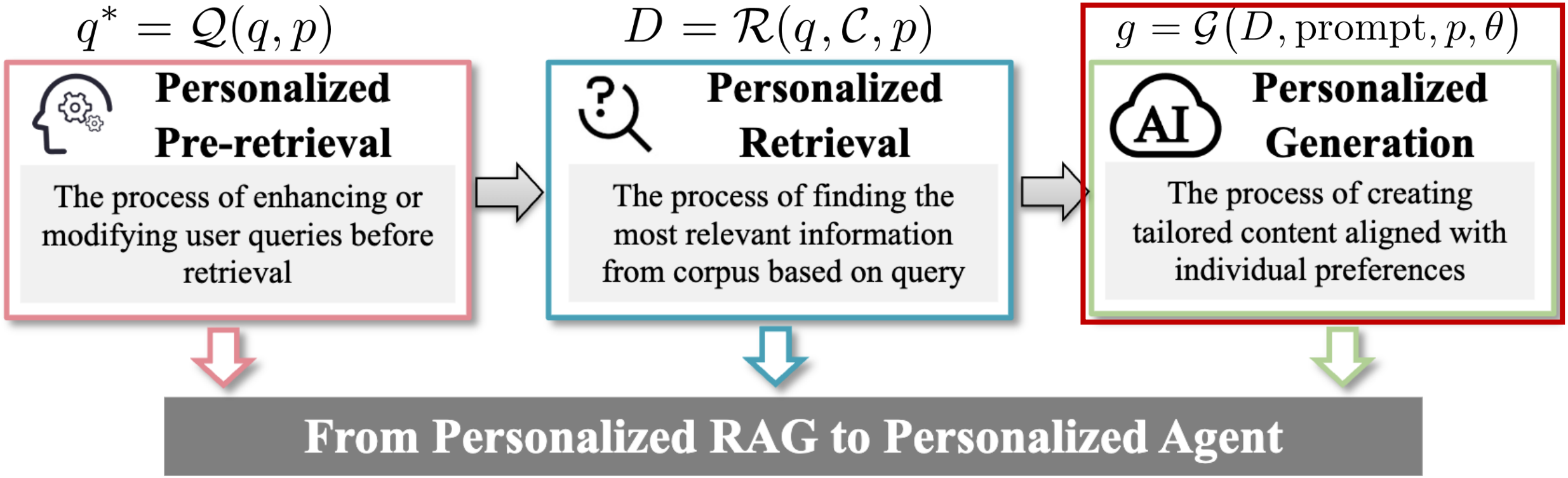


Proxy of the whole work flow.

$$g = \mathcal{G}\left(\mathcal{R}\left(\mathcal{Q}\left(q, p\right), \mathcal{C}, p\right), \text{prompt}, p, \theta\right)$$

Data & Evaluation Methodology

➢ Data & Evaluation Methodology in different stages:



Proxy of the whole work flow.

$$g = \mathcal{G}\left(\mathcal{R}\left(\mathcal{Q}\left(q, p\right), \mathcal{C}, p\right), \text{prompt}, p, \theta\right)$$

Data & Evaluation Methodology

➢ Data & Evaluation Methodology in different stages:

$q^* = \mathcal{Q}(q, p)$

Personalized Pre-retrieval

The process of enhancing or modifying user queries before retrieval

$D = \mathcal{R}(q, \mathcal{C}, p)$

Personalized Retrieval

The process of finding the most relevant information from corpus based on query

$g = \mathcal{G}(D, \text{prompt}, p, \theta)$

Personalized Generation

The process of creating tailored content aligned with individual preferences

From Personalized RAG to Personalized Agent

Proxy of the whole work flow.

$$g = \mathcal{G}\left(\mathcal{R}\left(\mathcal{Q}\left(q, p\right), \mathcal{C}, p\right), \text{prompt}, p, \theta\right)$$

Pre-retrieval Stage

➢Pre-retrieval Stage: Query Rewriting & Query Expansion

$$q^* = \mathcal{Q}(q, p)$$

Table 1. Summary of Dataset in Pre-retrieval Stage

Field	Dataset	Domain/Source	Personalized Signal	Metrics
Query Rewriting	SCAN [75, 226]	Synthetic commands	Weak (lack personalized feature)	Accuracy
	Robust04 [82]	News corpus retrieval	Medium (no direct personalized feature)	Accuracy
	Avocado Email [76, 118]	Enterprise email threads	Strong	BLEU, ROUGE
	Amazon Review [76, 116]	E-commerce reviews(no direct personalized feature)	Medium (no direct personalized feature)	BLEU, ROUGE
	Reddit Comments [76, 155]	Social comments	Medium (no direct personalized feature)	BLEU, ROUGE
	Amazon ESCI [115, 131]	Shopping query-product relevance	Medium (no direct personalized feature)	EM, ROUGE-L, XEntropy
	AOL Query Log [121, 227]	Web search query log	Strong	MAP, MRR, P@1
	WARRIORS [133]	LLM-simulated Web sessions	Strong	MRR, NDCG
	AITA & WORKSM[71, 112]	Reddit + workplace posts	Strong	Macro-F1, BS-F1
	PIP	Text-to-image prompts	Strong	PMS, Image-Align, ROUGE-L
Query Expansion	PR2 [9]	Personalized re-ranking	Strong	MAP, MRR, NDCG, RBP
	Delicious [12, 18]	Social bookmarking	Strong	Precision, MAP, MRR, Recall
	Flickr [12]	Photo tagging	Strong	Precision
	CiteULike [13, 17]	Academic bookmarking	Strong	Recall, MAP, MRR
	LRDP [15]	Media preferences	Medium (no direct personalized feature)	Precision, Recall, F1
	BibSonomy [111]	Literature tags	Strong	MAP, Precision, PQEC, $\text{Prof}_{\text{overlap}}$
	Wikipedia [11, 43]	Knowledge graph	Strong	Precision, Recall

Pre-retrieval Stage

➢Pre-retrieval Stage: Query Rewriting & Query Expansion

$$q^* = \mathcal{Q}(q, p)$$

Table 1. Summary of Dataset in Pre-retrieval Stage

Field	Dataset	Domain/Source	Personalized Signal	Metrics
Query Rewriting	SCAN [75, 226]	Synthetic commands	Weak (lack personalized feature)	Accuracy
	Robust04 [82]	News corpus retrieval	Medium (no direct personalized feature)	Accuracy
	Avocado Email [76, 118]	Enterprise email threads	Strong	BLEU, ROUGE
	Amazon Review [76, 116]	E-commerce reviews(no direct personalized feature)	Medium (no direct personalized feature)	BLEU, ROUGE
	Reddit Comments [76, 155]	Social comments	Medium (no direct personalized feature)	BLEU, ROUGE
	Amazon ESCI [115, 131]	Shopping query-product relevance	Medium (no direct personalized feature)	EM, ROUGE-L, XEntropy
	AOL Query Log [121, 227]	Web search query log	Strong	MAP, MRR, P@1
	WARRIORS [133]	LLM-simulated Web sessions	Strong	MRR, NDCG
	AITA & WORKSM[71, 112]	Reddit + workplace posts	Strong	Macro-F1, BS-F1
	PIP	Text-to-image prompts	Strong	PMS, Image-Align, ROUGE-L
Query Expansion	PR2 [9]	Personalized re-ranking	Strong	MAP, MRR, NDCG, RBP
	Delicious [12, 18]	Social bookmarking	Strong	Precision, MAP, MRR, Recall
	Flickr [12]	Photo tagging	Strong	Precision
	CiteULike [13, 17]	Academic bookmarking	Strong	Recall, MAP, MRR
	LRDP [15]	Media preferences	Medium (no direct personalized feature)	Precision, Recall, F1
	BibSonomy [111]	Literature tags	Strong	MAP, Precision, PQEC, $\text{Prof}_{\text{overlap}}$
	Wikipedia [11, 43]	Knowledge graph	Strong	Precision, Recall

Pre-retrieval Stage

- Avocado Email (Query Rewriting):
  - Domain/Source: **Enterprise email threads**
  - Personalized Signal: **Strong**

Table 1. Statistics of Avocado Email

Type	Total	Redact	Dup	Final	No Text	Text
Email	938,035	0	323,574	614,461	0	614,461
Attachment	325,506	110,023	126,411	89,072	23,731	65,341
Extracted file	556,286	298,022	172,870	85,394	14,775	70,619
Contact	59,455	0	0	59,455	59,455	0
Appointment	76,902	0	0	76,902	0	76,902
Task	15,474	0	0	15,474	0	15,474
Schedule	26,980	0	0	26,980	0	26,980
Report	7,206	0	0	7,206	7,206	0
Journal	4,204	4,204	0	0	0	0
Stickynote	2,232	2,232	0	0	0	0
Other	21,460	21,460	0	0	0	0
Total	**2033740**	**435941**	**622855**	**974944**	**105167**	**869777**

Pre-retrieval Stage

- Avocado Email (Query Rewriting):

Example Record (field names only; values illustrative)

The Email record has the following attributes: messageid is <Message-ID as in PST>; subject is RE: Q4 deployment checklist; in_reply_to is <parent Message-ID>; sentto_address is team-list@avocadoit.example; sent_date is 2003-11-14T15:22:01Z.

Figure 1. Example of Avocado Email

- Task framing: Current email thread + user history → rewritten query or personalized prompt for retrieval from internal mail store
- Metrics: **BLEU, ROUGE** (lexical overlap with reference rewrites)

Pre-retrieval Stage

- Amazon ESCI (Query Rewriting):
  - Domain/Source: **Shopping query-product relevance**
  - Personalized Signal: **Medium**

- **Exact (E):** the item is relevant for the query, and satisfies all the query specifications (e.g., a water bottle matching all attributes of a query "plastic water bottle 24oz", such as material and size)
- **Substitute (S):** the item is somewhat relevant, i.e., it fails to fulfill some aspects of the query but the item can be used as a functional substitute (e.g., fleece for a "sweater" query)
- **Complement (C):** the item does not fulfill the query, but could be used in combination with an exact item (e.g., track pants for "running shoes" query)
- **Irrelevant (I):** the item is irrelevant, or it fails to fulfill a central aspect of the query (e.g., socks for a "telescope" query, or a wheat flour bread for a "gluten–free bread" query)

Pre-retrieval Stage

- Amazon ESCI (Query Rewriting):

Table 1: Summary of the Shopping queries dataset for the task 1 (small version): the number of unique queries, the number of judgements, and the average number of judgements per query (Avg. Depth).

	Total			Train			Public Test		
Language	# Queries	# Judgements	Avg. Depth	# Queries	# Judgements	Avg. Depth	# Queries	# Judgements	Avg. Depth
English (US)	29,844	601,462	20.2	20,888	419,730	20.1	4,477	91,062	20.3
Spanish (ES)	8,049	218,826	27.2	5,632	152,917	27.2	1,208	32,905	27.2
Japanese (JP)	10,407	297,882	28.6	7,284	209,091	28.7	1,561	43,832	28.1
Overall	48,300	1,118,170	23.2	33,804	781,738	23.1	7,246	167,799	23.2

Table 2: Summary of the Shopping queries dataset for the tasks 2 and 3 (large version): the number of unique queries, the number of judgements, and the average number of judgements per query (Avg. Depth).

	Total			Train			Public Test		
Language	# Queries	# Judgements	Avg. Depth	# Queries	# Judgements	Avg. Depth	# Queries	# Judgements	Avg. Depth
English (US)	97,345	1,819,105	18.7	68,139	1,272,626	18.7	14,602	274,261	18.8
Spanish (ES)	15,180	356,578	23.5	10,624	249,721	23.5	2,277	53,494	23.5
Japanese (JP)	18,127	446,055	24.6	12,687	312,397	24.6	2,719	66,612	24.5
Overall	130,652	2,621,738	20.1	91,450	1,834,744	20.1	19,598	394,367	20.1

- Task framing: Task 1: Query-Product Ranking; Task 2: Product Classification; Task 3: Product Substitute Identification
- Metrics: **NDCG,EM, ROUGE-L, X-Entropy**

Pre-retrieval Stage

- Amazon ESCI (Query Rewriting):

Table 1: Summary of the Shopping queries dataset for the task 1 (small version): the number of unique queries, the number of judgements, and the average number of judgements per query (Avg. Depth).

	Total			Train			Public Test		
Language	**# Queries**	**# Judgements**	**Avg. Depth**	**# Queries**	**# Judgements**	**Avg. Depth**	**# Queries**	**# Judgements**	**Avg. Depth**
English (US)	29,844	601,462	20.2	20,888	419,730	20.1	4,477	91,062	20.3
Spanish (ES)	8,049	218,826	27.2	5,632	152,917	27.2	1,208	32,905	27.2
Japanese (JP)	10,407	297,882	28.6	7,284	209,091	28.7	1,561	43,832	28.1
Overall	48,300	1,118,170	23.2	33,804	781,738	23.1	7,246	167,799	23.2

Table 2: Summary of the Shopping queries dataset for the tasks 2 and 3 (large version): the number of unique queries, the number of judgements, and the average number of judgements per query (Avg. Depth).

	Total			Train			Public Test		
Language	**# Queries**	**# Judgements**	**Avg. Depth**	**# Queries**	**# Judgements**	**Avg. Depth**	**# Queries**	**# Judgements**	**Avg. Depth**
English (US)	97,345	1,819,105	18.7	68,139	1,272,626	18.7	14,602	274,261	18.8
Spanish (ES)	15,180	356,578	23.5	10,624	249,721	23.5	2,277	53,494	23.5
Japanese (JP)	18,127	446,055	24.6	12,687	312,397	24.6	2,719	66,612	24.5
Overall	130,652	2,621,738	20.1	91,450	1,834,744	20.1	19,598	394,367	20.1

- Task framing: Task 1: Query-Product Ranking; Task 2: Product Classification; Task 3: Product Substitute Identification
- Metrics: **NDCG,EM, ROUGE-L, X-Entropy**

Pre-retrieval Stage

- AOL Query Log (Query Rewriting):
  - Domain/Source: **Web search query log**
  - Personalized Signal: **Strong**

Table 1. Statistics of AOL Query Log

Metric Description	Count
Lines of data	36,389,567
Instances of new queries (w/ or w/o click-through)	21,011,340
Requests for "next page" of results	7,887,022
User click-through events	19,442,629
Queries w/o user click-through	16,946,938
Unique (normalized) queries	10,154,742
Unique user IDs	657,426

Pre-retrieval Stage

- AOL Query Log (Query Rewriting):

Example Record (field names only; values illustrative)

The AOL Query Log session has the following sequential events:
Event 1 (Query Only): AnonID is 35412; Query is avocado health benefits; QueryTime is 2006-03-15T10:22:01Z.
Event 2 (Click-through): AnonID is 35412; Query is avocado health benefits; QueryTime is 2006-03-15T10:22:15Z; ItemRank is 2; ClickURL is medicalnewstoday.com.
Event 3 (Next Page): AnonID is 35412; Query is avocado health benefits; QueryTime is 2006-03-15T10:23:00Z.

Figure 1. Example of AOL Query Log

- Task framing: Current query + user session history (previous queries & clicked domains) -> rewritten (contextualized/expanded) query for personalized web search retrieval
- Metrics: **MAP, MRR, P@1**

Pre-retrieval Stage

- Wikipedia (Query Expansion):
  - Domain/Source: **Knowledge Graph**
  - Personalized Signal: **Strong**

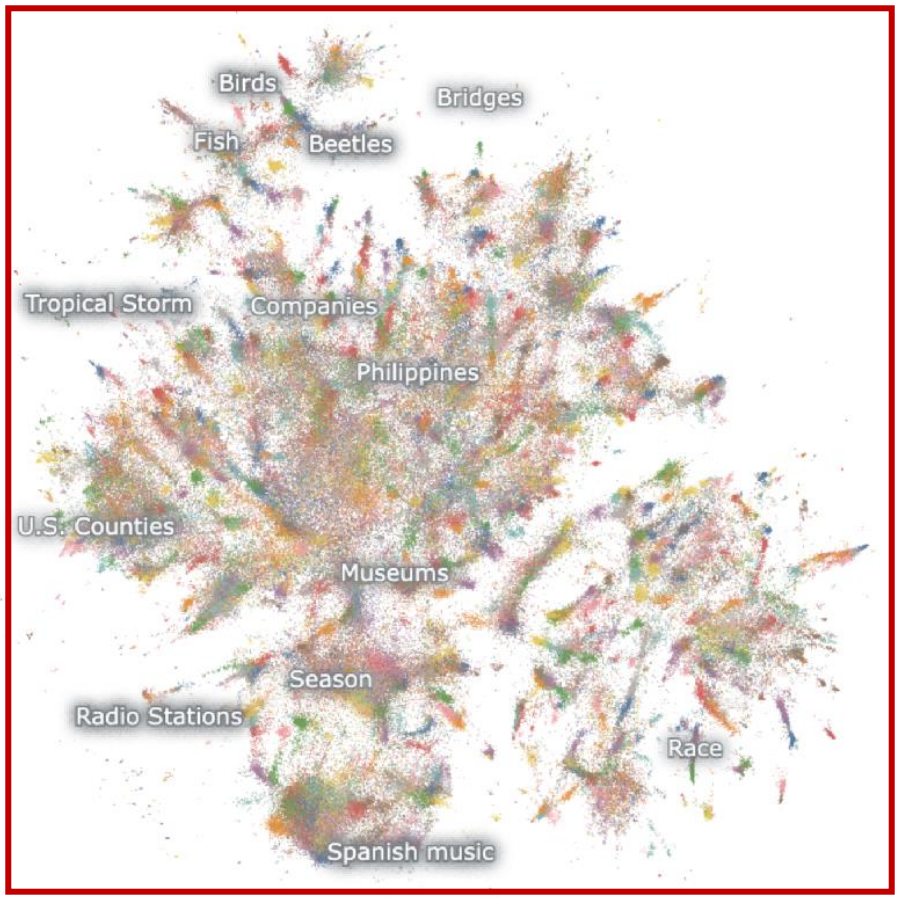


Figure 1. Visualization of Wikipedia

Table 1. Statistics of Wikipedia

Attribute	Details
Overall Statistics	
Total Number of Subsets	323 (Languages)
Overall Size	10M – 100M rows
Data Splits	`train` only
Format	Parquet
Data Fields (Features)	
`id`	Article ID (String)
`url`	Article URL (String)
`title`	Article Title (String)
`text`	Article Content (String)
Top Languages by Row Count (Snapshot: 20231101)	
English (en)	~ 6.41M
Cebuano (ceb)	~ 6.12M
German (de)	~ 2.85M
French (fr)	~ 2.56M
Spanish (es)	~ 1.84M
Chinese (zh)	~ 1.38M

Pre-retrieval Stage

- Wikipedia (Query Expansion):
  - Domain/Source: **Knowledge Graph**
  - Personalized Signal: **Strong**

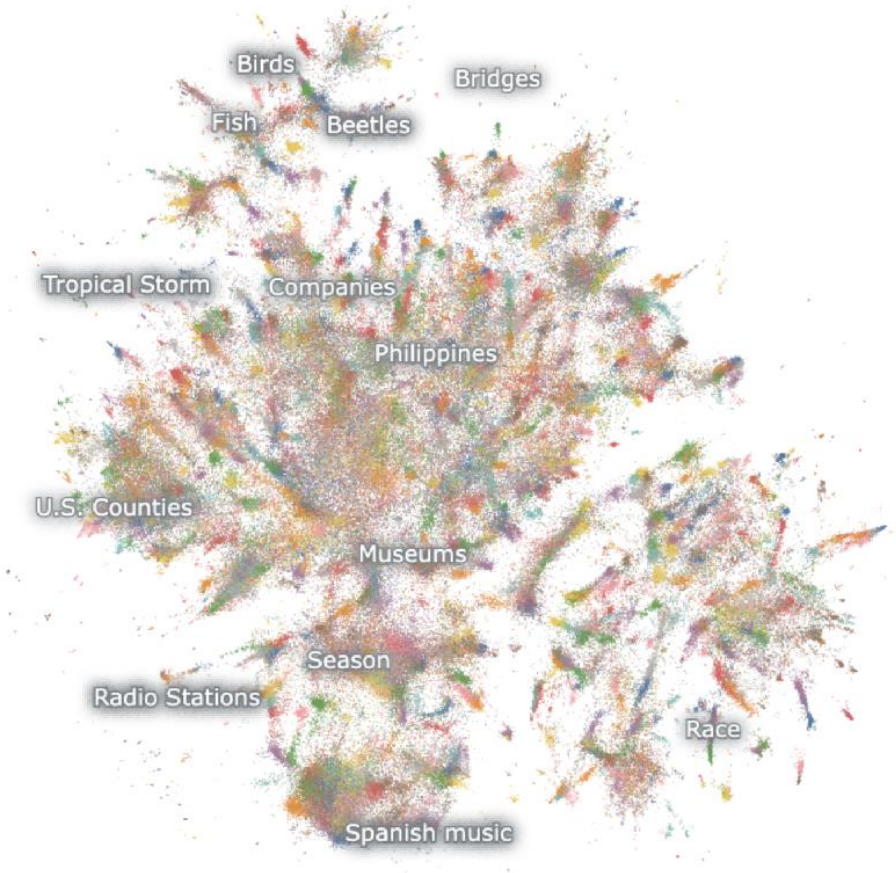


Figure 1. Visualization of Wikipedia

Table 1. Statistics of Wikipedia

Attribute	Details
Overall Statistics	
Total Number of Subsets	323 (Languages)
Overall Size	10M – 100M rows
Data Splits	`train` only
Format	Parquet
Data Fields (Features)	
`id`	Article ID (String)
`url`	Article URL (String)
`title`	Article Title (String)
`text`	Article Content (String)
Top Languages by Row Count (Snapshot: 20231101)	
English (en)	~ 6.41M
Cebuano (ceb)	~ 6.12M
German (de)	~ 2.85M
French (fr)	~ 2.56M
Spanish (es)	~ 1.84M
Chinese (zh)	~ 1.38M

Pre-retrieval Stage

- Wikipedia (Query Expansion):

Example Record (field names only; values illustrative)

The Wikipedia record has the following attributes:
id is 1; url is https://simple.wikipedia.org/wiki/April;
title is April;
text is April is the fourth month of the year in the Julian and Gregorian calendars, and comes between March and May....

Figure 1. Example of Wikipedia

- Task framing: Current query + Retrieved Wikipedia context (e.g., related title or keywords from text) -> Expanded query enriched with semantic concepts for retrieval
- Metrics: **Precision, Recall**

Retrieval Stage

➢ Retrieval Stage: $D = \mathcal{R}(q, \mathcal{C}, p)$

Table 1. Summary of Dataset in Retrieval Stage

Field	Dataset	Domain / Source	Personalized Signal	Metrics
Retrieval	TOPDIAL [167]	Synthetic role-play	Strong	BLEU, P-F1, K-F1, Succ. (%)
	LiveChat [44]	Chats History	Strong	Recall@k, MRR
	PersonalityEvd [156]	Chinese TV series	Medium (Scripted characters, not authentic)	Accuracy, Human Eval
	Pchatbot [127]	Weibo logs	Strong	BLEU, ROUGE, MRR
	DuLeMon [192]	Chat History	Strong	PPL, BLEU Acc, Precision, Recall, F1
	PersonalityEdit [104]	Synthetic opinion QA	Strong	ES, DD, Acc

Retrieval Stage

- Pchatbot:
  - Domain/Source: **Weibo logs**
  - Personalized Signal: **Strong**

Table 1. Statistics of PChatbot

Dataset	# Dialogues	# Utterances	# Words	Language	Source	Personalized Info.
PchatbotW (Our)	139,448,339	278,896,678	8,512,945,238	Chinese	Weibo	User ID & Timestamp
PchatbotL (Our)	59,427,457	118,854,914	3,031,617,497	Chinese	Judicial Forums	User ID & Timestamp

	PchatbotW-R	PchatbotW-G
# Users	420,000	300,000
Avg. history length	32.3	11.4
Avg. # words of post	24.9	22.9
Avg. # words of response	10.1	9.6
# Response candidates	10	N/A
# Training samples	3,000,000	2,707,880
# Validation samples	600,000	600,000
# Testing samples	600,000	600,000

	PchatbotW	PchatbotL	PchatbotW-1	PchatbotL-1
# Posts	5,319,596	20,145,956	3,597,407	4,662,911
# Responses	139,448,339	59,427,457	13,992,870	5,523,160
# Users in posts	772,002	5,203,345	417,294	1,107,989
# Users in resp.	23,408,367	203,636	2,340,837	20,364
Avg. # resp. / post	26.214	2.950	3.890	1.184
Max. # resp. / post	525	120	136	26
Vocabulary Size	9,148,532	1,329,930	3,447,433	551,071
Avg. # words of post	49.37	36.88	49.40	37.26
Avg. # words of resp.	11.68	14.08	11.70	14.13

Retrieval Stage

- Pchatbot:

Table 3: An example of data in Pchatbot

Field	Content
Post	下冰雹了!真刺激! (Hailing! It's really exciting!)
Post user ID	5821954
Post timestamp	634760927
Response	出去感受更刺激 (It's more exciting to go out.)
Response user ID	592445
Response timestamp	634812525
Partition index (1-10)	1
Train/Dev/Test (0/1/2)	0

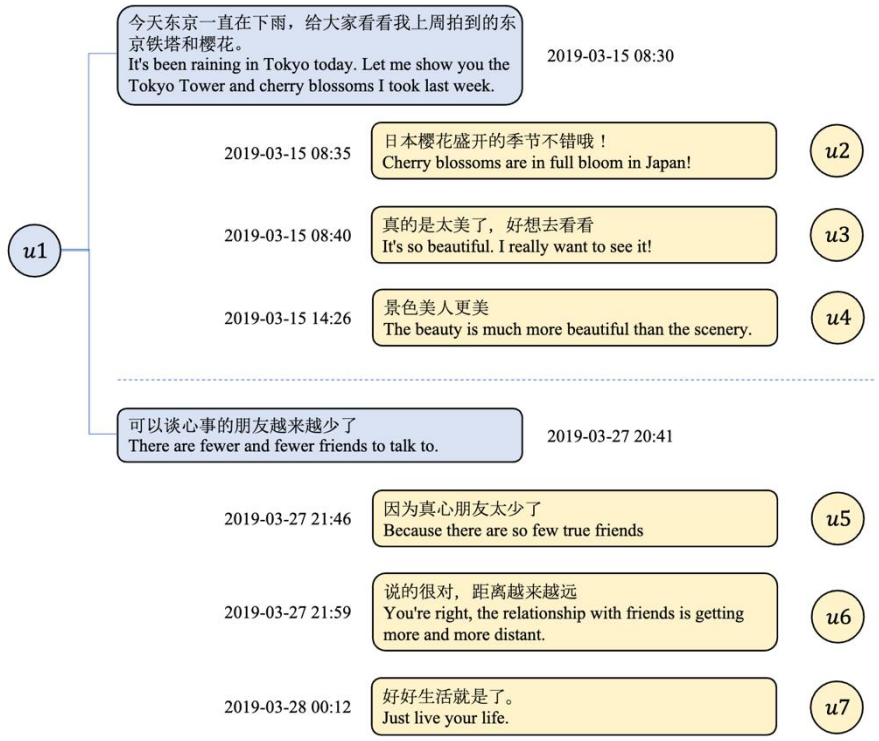


Figure 1. Example of Pchatbot

- Task framing: current post as query + candidate response corpus + target user history -> retrieve and rank responses
- Metrics: **BLEU, ROUGE, MRR**

Retrieval Stage

- DuLeMon:
  - Domain/Source: **Chat history**
  - Personalized Signal: **Strong**

Table 1. Statistics of DuLeMon

Dataset	Persona	Mutual	# Dialogues	Language	Multi-turn
DuLeMon (Ours)	Text	✓	27,501	Chinese	Yes

Category	SELF	BOTH
# Dialogues	24500	3001
# Utterances	400472	48522
Avg. # turns	16.3	16.2
Avg. length of utterances	19.7	21.2
Avg. # bot persona	4.0	4.0
Avg. # user persona (seen)	0	4.4
Avg. # user persona (unseen)	4.0	1.3

Retrieval Stage

➢ DuLeMon:

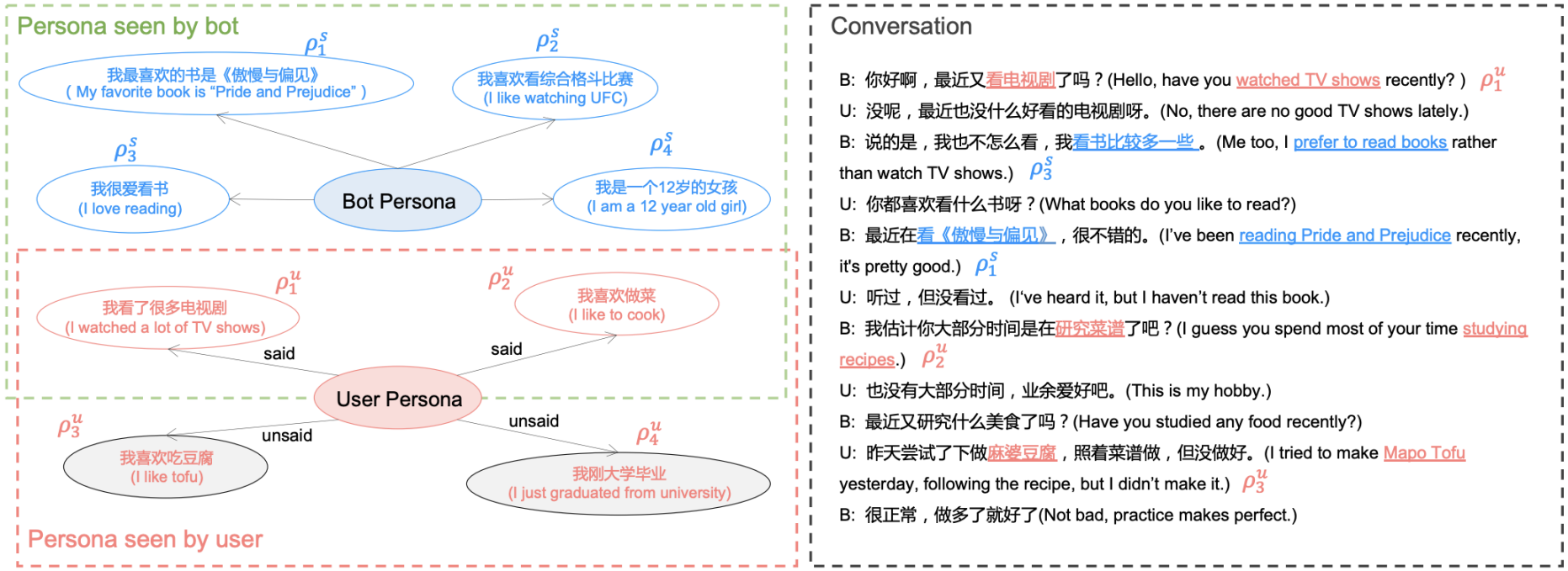


Figure 1. Example of DuLeMon

➢ Task framing: current dialogue context + long-term user and bot persona memories -> retrieve memories and generate response

➢ Metrics: **PPL, BLEU, Acc, Precision, Recall, F1**

Generation Stage

➢Generation Stage: $g = \mathcal{G}(D, \text{prompt}, p, \theta)$

Table 1. Summary of Dataset in Generation Stage

Field	Benchmark / Dataset	Domain / Source	Personalized Signal	Metrics
Generation	LaMP [139]	Multi-domain personalized task	Strong	Accuracy, F1, MAE, RMSE, ROUGE
	LongLaMP [74]	Long-horizon personalized tasks	Strong	Accuracy, F1, MAE, RMSE, ROUGE
	DuLeMon [192]	Open-domain dialogue	Strong	PPL, BLEU, DISTINCT, F1
	PGraphRAG [8]	E-commerce interaction graphs	Strong	ROUGE, METEOR, MAE, RMSE
	AmazonQA/Products [36]	Product QA	Strong	ROUGE, Persona F1
	LoCoMo [102]	conversations	Strong	QA F1, R@k, FactScore/ROUGE-L, NLG metrics

➢Generation evaluation asks whether the final answer is not only fluent and faithful, but aligned with a specific user's writing style, preferences, and historical behavior

Generation Stage

- LaMP:

Task	Input	User Profile	Target Output
LaMP-1 Citation Identification	paper information + two candidate references	author's historical papers, usually title + abstract	personalized citation choice
LaMP-2 Movie Tagging	movie description	user's historical movie-tag pairs	personalized movie tag
LaMP-3 Product Rating	product review	user's historical review-rating pairs	personalized product rating
LaMP-4 News Headline	news article	author's historical article-title pairs	personalized headline
LaMP-5 Scholarly Title	paper abstract	author's historical abstract-title pairs	personalized paper title
LaMP-6 Email Subject	email body	sender's historical email-subject pairs	personalized subject
LaMP-7 Tweet Paraphrasing	tweet text	user's historical tweets	tweet in user style

- Benchmark Role: evaluate whether an LM can produce outputs conditioned on user-specific historical data rather than using a one-size-fits-all prompt
- Metrics: **ROUGE-1, ROUGE-L**
- Evaluation settings: user-based split for new users and time-based split for future interactions of existing users

Generation Stage

- LaMP:

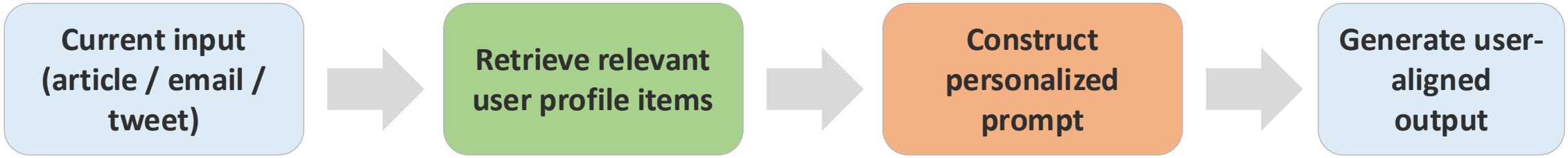


Figure 1. LaMP generation framing: input + retrieved profile → personalized output

- Task framing : current input + historical user profile → produce a personalized target output
- Retrieval augmentation is central: the system selects profile entries with BM25, Contriever, Recency, Random, or tuned retrieval strategies before generation

Generation Stage

- LongLaMP:

Task	Input	User Profile	Target Output
Email Completion	subject + partial email	previous subject-email pairs	long email continuation
Abstract Generation	paper title + keywords	previous authored papers	scientific abstract
Review Writing	product description + rating + summary	previous reviews and ratings	long product review
Topic Writing	Reddit post summary	previous summary-content pairs	full Reddit post

- Domain/Source: Avocado Email, Citation Network, Amazon Reviews, Reddit TL;DR
- Benchmark role: personalized long-text generation, where outputs span multiple sentences or paragraphs rather than short labels or titles
- Metrics: **ROUGE-1, ROUGE-L, METEOR**
- Settings: user split for cold-start users and temporal split for known users with evolving writing style

Generation Stage

- LongLaMP:

Table 1. LongLaMP statistics (selected mean values)

Task	Setting	#Train	#Val	#Test	Avg. Output Len.	Avg. Profile Size
Email Completion	User	3,286	958	823	92.6	85.7
Email Completion	Temporal	3,234	833	818	92.8	58.9
Abstract Generation	User	13,693	4,560	4,560	144.3	120.3
Abstract Generation	Temporal	22,822	4,565	4,564	160.7	118.2
Review Writing	User	14,745	1,826	1,822	304.5	34.4
Topic Writing	User	11,442	2,452	2,453	263.0	50.4

- Task framing: input prompt + retrieved user profile → generate a long personalized target text
- Evaluation pressure is higher than LaMP: the model must preserve user style, keep coherence over longer spans, and avoid topic drift

Agent Stage

➢Agent: Role-Playing Agent Data & Evaluation

Table 1. Summary of Dataset in Agent Flow

Field	Dataset name	Agent Usage	Dataset Source	Application Area
Agent	MIT-BIH Arrhythmia [190]	Physiological signal understanding; real-world sensing demos	PhysioNet + self-collected	Healthcare signals / physical world
	Abbasian et al. [2]	Personalized health dialogue agent	–	Healthcare dialogue
	RoleBench [179]	Train/evaluate role-playing ability	Author-released	Role-playing dialogue
	character-llm data [145]	Train agents with character experiences/memory	Author-released	Role-playing dialogue
	InCharacter [173]	Interview dialogues for personality fidelity	Author-released	Personality evaluation
	SocialBench [22]	Evaluate sociality/interaction skills	Author-released	Social conversational ability
	MMRole-Data [34]	Train/evaluate text–image role agents	Author-released	Multimodal role-playing
	RolePersonality [130]	Personalized dialog finetuning	Author-released	Role-playing
	Wang et al. [163]	Conversational control over mobile UI	–	Mobile UI manipulation
	CharacterEval [160]	Evaluate Chinese role-playing agents	Author-released	Chinese role-playing dialogue
	Yu et al. [198]	Efficient multi-character adaptation	–	Role-playing dialogue
	Wang et al. [177]	Personalization memory graphs	Private	Personal assistant / enterprise
	Smallville [120]	Agent memory/behavior evaluation	Author-released	Social simulation / HCI
	Sim-User [2]	Simulate user behavior/conversations	Author-released	User simulation
	Singh et al. [150]	Personalized itinerary planning	–	Travel planning
	PersonalWAB [20]	Train/evaluate personalized web agents	Author-released	Web task agents
	Minecraft Wang et al. [164]	Open-ended skill exploration	Open environment	Embodied games
	ALFRED environment [205]	LLM-guided skill transfer	Public benchmarks (varied)	Problem solving / skill transfer
	Huang et al. [57]	High-level tasks plans	Public	Embodied task agent
	Zheng et al. [223]	Goal-driven multi-agent collaboration	–	Organizational/management agents
	KBP [51]	Persona-grounded, source-aware dialogue	Author-released	Personalized knowledge dialogue
	Renze and Guven [134]	Compare self-reflection strategies for solving problems	Author-released	Reasoning / problem solving
	Wang et al. [174]	Personality consistency Evaluation	–	Role-playing / personality
	LoCoMo [102]	Evaluate long-term conversational memory	Author-released	Long-term conversational memory

Agent Stage

➢Data & Evaluation Methodology in agent flow:

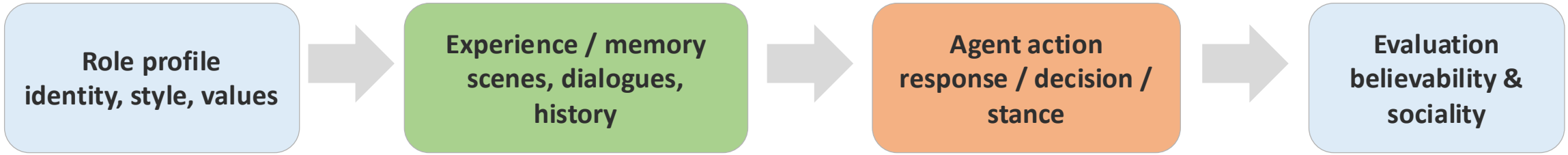


Figure 1. Agent evaluation flow: persistent identity + memory → interactive behavior → social evaluation

Stage	What is evaluated	Example signal
Identity consistency	Does the agent preserve role knowledge, speaking style, and values?	Character profile, role-specific knowledge
Memory behavior	Does the agent recall prior experience or conversation context?	Experience scenes, long dialogue context
Social behavior	Does the agent respond appropriately in individual and group interactions?	Emotion, preference, cooperation/conflict dynamics

➢This flow connects personalized RAG with agents: retrieved memory and role profile are no longer only evidence for one answer; they become persistent state used across turns

Agent Stage

➢Character-LLM Data:

Table 1. Selected statistics of Character-LLM experience data

Character	Scenes	Words	Turns / Scene	Words / Turn
Cleopatra VII	1.4K	723K	14.3	36
Julius Caesar	1.6K	820K	12.9	39
Ludwig van Beethoven	1.6K	663K	12.2	33
Socrates	1.6K	896K	14.1	41
Martin Luther King	2.2K	1,038K	12.0	40
Average	1.6K	754K	13.2	36

➢Domain/Source: public character profiles; historical figures, fictional characters, and celebrities

➢Core idea: instead of only prompting a generic LLM to act as a role, Character-LLM fine-tunes a separate agent on reconstructed character experiences

Agent Stage

➢Character-LLM Data:

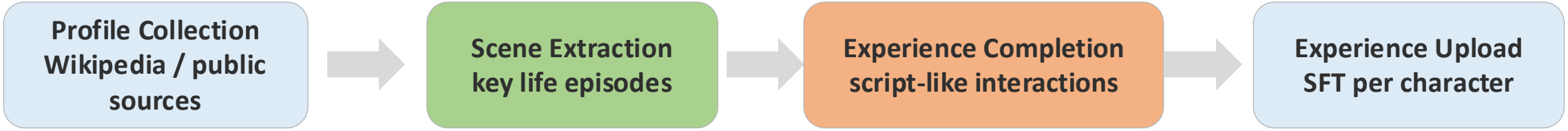


Figure 1. Character-LLM construction flow

Dimension	What it checks
Memorization	recall of character-specific people, events, objects, and experience
Values / Personality	whether stance, objectives, tone, and emotional reaction match the role
Hallucination / Stability	whether the agent avoids out-of-role knowledge and remains consistent over long acting

➢Task framing: character profile + generated scenes → train a role-playing agent → interview the agent and judge whether it preserves character identity

➢Evaluation setting: single-turn and multi-turn interviews; 1,307 collected interview questions across nine characters

Agent Stage

- SocialBench / RoleInteract:

Table 1. Evaluation dimensions in SocialBench / RoleInteract

Level	Dimension	Subcategories
Individual	Self-Awareness	role style; role knowledge
Individual	Emotional Perception	situational understanding; emotion detection
Individual	Conversation Memory	short-term memory; long-term memory
Group	Social Preference	positive; neutral; negative group dynamics

- Domain/Source: books, movies, novels, online role-playing platforms, and LLM-generated role-play dialogues
- Benchmark role: evaluate sociality of role-playing conversational agents at both individual and group levels

Agent Stage

- SocialBench / RoleInteract:

Table 1. Statistics by Category in SocialBench / RoleInteract

Category	SA Style	SA Know.	EP Situ.	EP Emo.	CM Short	CM Long	Pos.	Neu.	Neg.
# Questions	1,063	1,408	193	1,016	773	1,348	586	724	606
Avg. Utterances	17.9	9.4	1.0	6.4	23.9	76.7	15.6	16.1	16.0
Avg. Tokens / Utt.	32.6	66.7	286.3	23.0	37.6	41.2	38.8	38.7	42.0

- Task framing: role profile + dialogue/group context + instruction → select an answer or generate a response that matches role identity and social setting
- Data construction: role profiles are collected from diverse sources; dialogues are extracted from platforms or generated via role-playing/self-dialogue; questions are manually validated
- Metrics: single-answer accuracy; multiple-answer partial-credit accuracy; keyword coverage rate for open-domain generation questions.

Individual: identity + emotion + memory

Group: social preference under multi-agent dynamics

Summary

➢Summary:

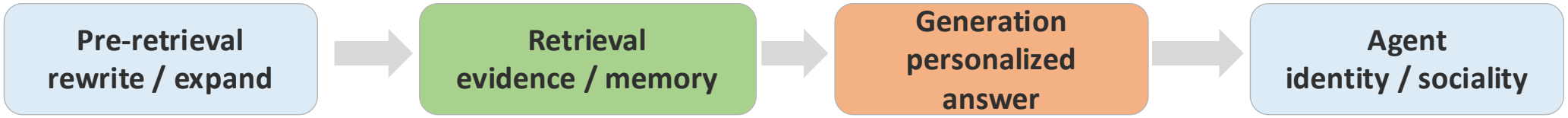


Figure 1. Evaluation unit shifts from query processing to long-term agent behavior

Stage	Core evaluation question	Representative datasets / benchmarks
Pre-retrieval	Does the query preserve and expose the user-specific intent?	Avocado Email, Amazon ESCI, AOL Query Log, Wikipedia
Retrieval	Does the system retrieve user-relevant evidence or memory?	PChatbot, DuLeMon
Generation	Is the final answer faithful, useful, and profile-aligned?	LaMP, LongLaMP
Agent	Does the agent maintain identity, memory, and social behavior over interaction?	Character-LLM data, SocialBench / RoleInteract

➢Key message: personalized RAG evaluation should be stage-wise; a single metric cannot cover query intent, retrieval quality, output personalization, and agent-level social consistency

Agenda

1. Introduction
9:00 – 9:50



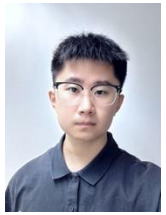
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

8. Hands on Projects
16:45 – 17:30



Pengyue Jia

Challenges

➢Six challenges that define the research frontier

01 Evaluation
How do we measure the degree and quality of personalization?

02 Privacy
How do we personalize without centralizing sensitive data?

03 Scalability
How do we serve many users without linear per-user cost?

04 Ethics & Fairness
How do we avoid bias and keep reasoning grounded?

05 Multimodality
How do we personalize across images, audio, video, and GUI traces?

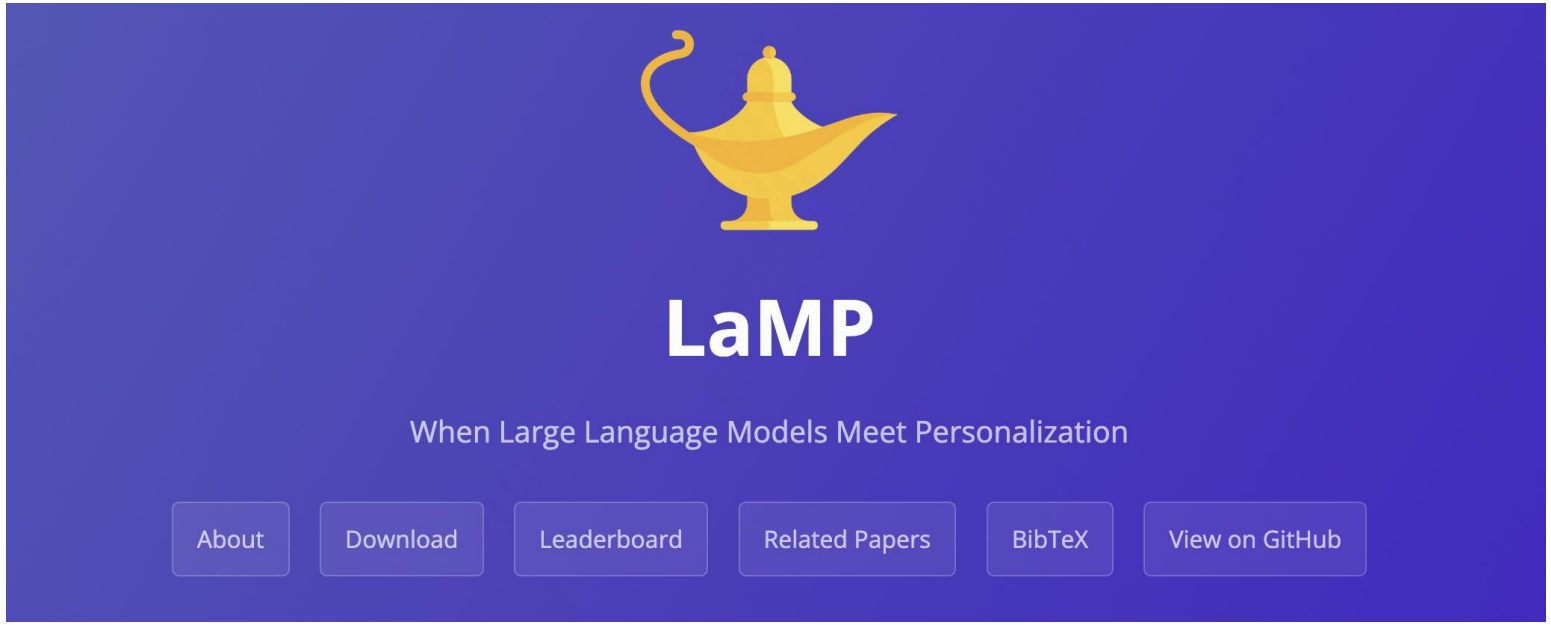
06 Agent Planning
How do we personalize decisions, tools, and long-horizon actions?

Challenges 1: Evaluation

- **Problem 1: Current metrics are insufficient**
  - BLEU/ROUGE measure surface overlap with a reference answer
  - In personalization, multiple answers can be correct depending on the user profile, preference, and interaction history
- **Problem 2: Human evaluation is costly**
  - Human evaluation is expensive and subjective
  - Hard to scale across users, domains, and longitudinal interactions
- **Problem 3: Open benchmark gap**
  - Current benchmarks often cover one stage
  - Personalized RAG requires holistic evaluation across pre-retrieval, retrieval, generation, and planning

Challenges 1: Evaluation

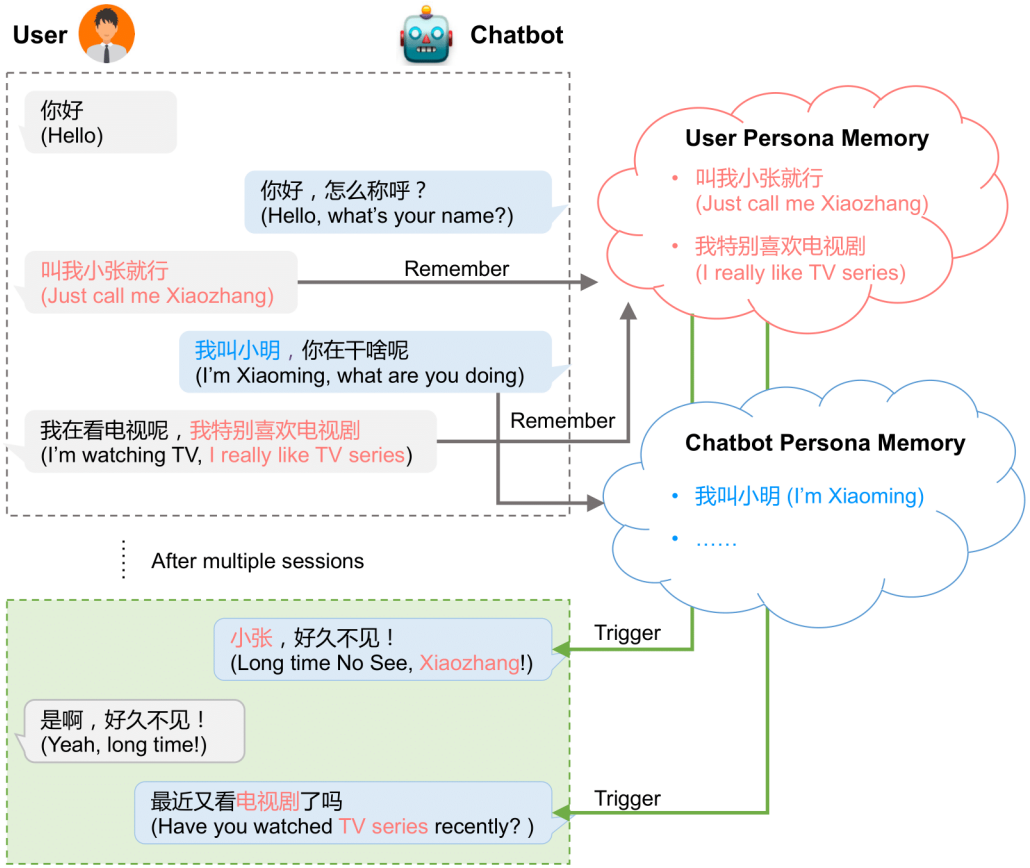
- **Representative works:**
  - **LaMP**
    - **Motivation**: Test whether a language model uses user profiles rather than producing generic outputs
    - **Contribution**: Tasks for personalized classification and generation
    - **Shortcomings**: Focuses mainly on output-level personalization, not the full pipeline



[1] Salemi, Alireza, et al. "Lamp: When large language models meet personalization." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024.

Challenges 1: Evaluation

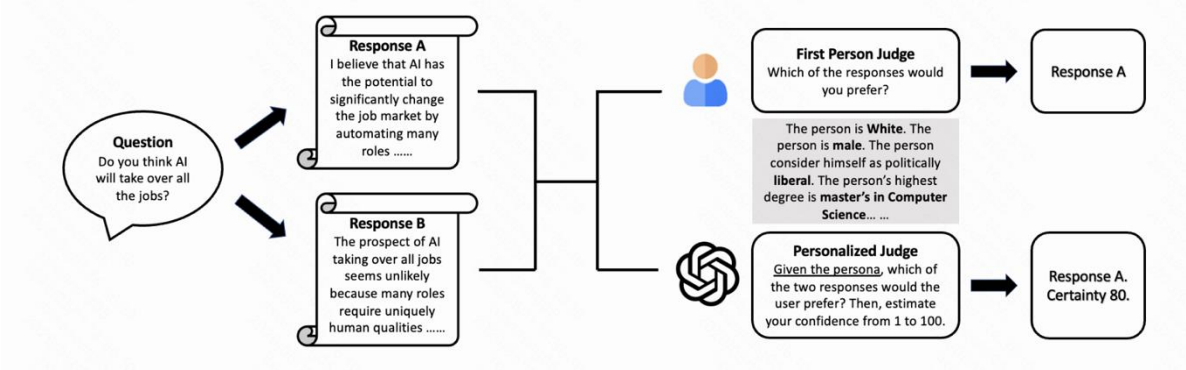
- **Representative works:**
  - **LaMP**
  - **DuLeMon**
    - **Motivation**: Personalization depends on retrieving the right user-relevant context
    - **Contribution**: Data for evaluating long-term personalized dialogue retrieval
    - **Shortcomings**: Focuses mainly on persona memory conversation retrieval



[1] Xu, Xinchao, et al. "Long time no see! open-domain conversation with long-term persona memory." Findings of the Association for Computational Linguistics: ACL 2022. 2022.

Challenges 1: Evaluation

- **Representative works:**
  - **LaMP**
  - **DuLeMon**
  - **LLM-as-Judge for personalization**
    - **Motivation**: human evaluation is expensive and hard to scale
    - **Contribution**: judge models can compare outputs for user alignment
    - **Shortcomings**: Lack multiple-aspect evaluation and user profile consistency



[1] Dong, Yijiang River, Tiancheng Hu, and Nigel Collier. "Can llm be a personalized judge?." Findings of the Association for Computational Linguistics: EMNLP 2024. 2024..

Challenges 1: Evaluation

- Representative works:
  - LaMP
  - DuLeMon
  - LLM-as-Judge for personalization

- What we need?
  - Measure for task correctness & profile alignment, satisfaction
  - Holistic evaluation across multiple stages
  - Multiple aspects & Extensive LLM-as-Judge

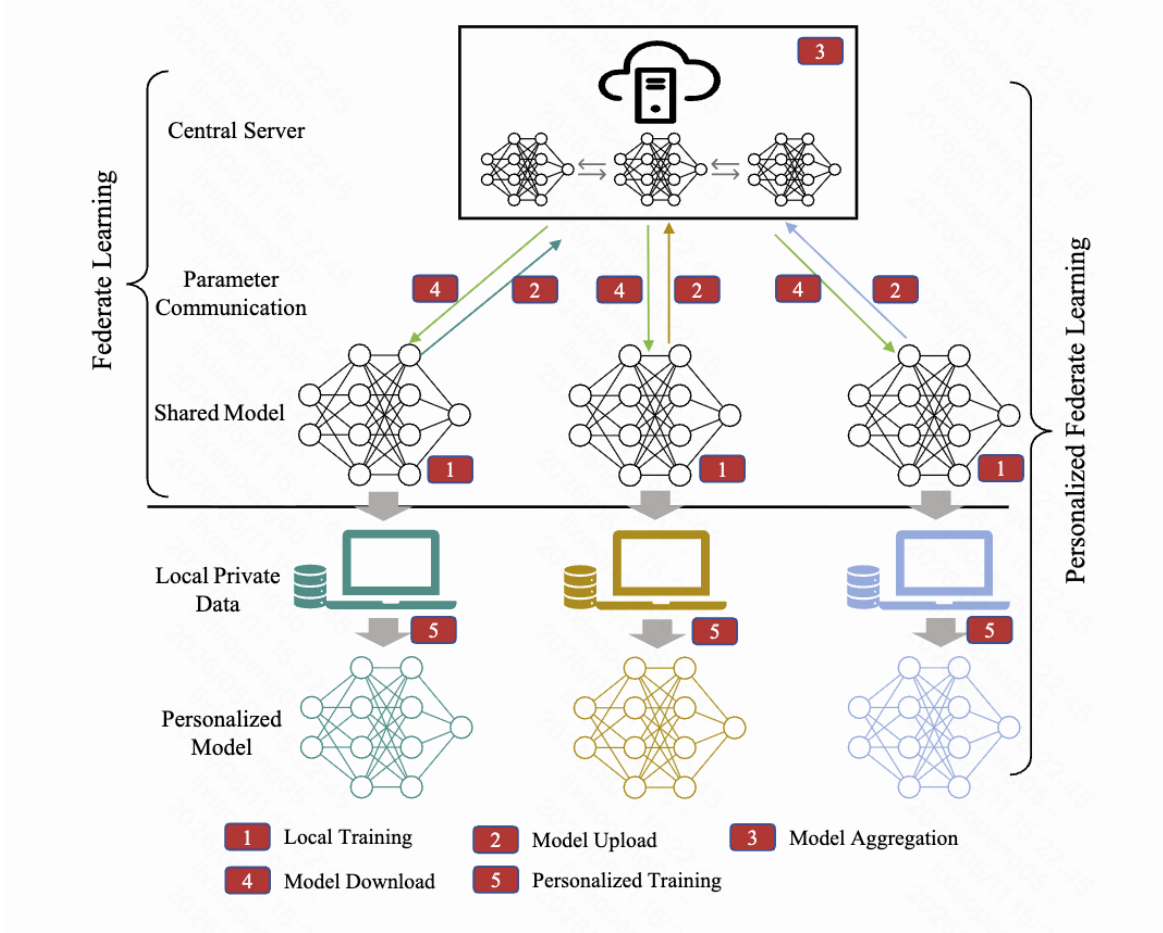
Challenges 2: Privacy

- **Problem 1: Sensitive data**
  - Search history
  - Location
  - Health information
  - Social context
  - Purchase intent, etc.

- **Problem 2: Centralized risk**
  - Sending raw queries and rich profiles to cloud LLMs can violate data locality
  - Increases breach risk, and undermine user trust

Challenges 2: Privacy

- **Representative works:**
  - **FDLoRA**
    - **Motivation**: Full fine-tuning or centralized training is too costly and exposes private user data
    - **Contribution**: Personalized federated learning with lightweight local adapter modules
    - **Method**: Train adapters locally; aggregate only privacy-preserving updates or insights



[1] Qi, Jiaxing, et al. "Fdlora: Personalized federated learning of large language model via dual lora tuning." arXiv preprint arXiv:2406.07925 (2024).

Challenges 2: Privacy

- **Representative works:**
  - **FDLoRA**

- **Potential future research area:**
  - **On-device personalization + cloud-scale reasoning**
    - Local model handles sensitive user-specific features
    - Cloud model receives constrained, abstracted context
    - Federated updates and encryption reduce leakage surface

[1] Qi, Jiaxing, et al. "Fdlora: Personalized federated learning of large language model via dual lora tuning." arXiv preprint arXiv:2406.07925 (2024).

Challenges 3: Scalability

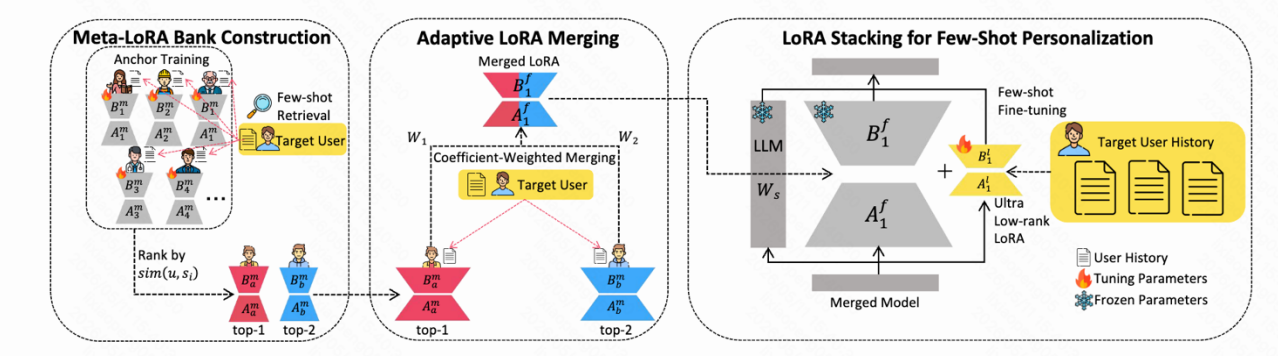
- **Problem 1: Retriever bottleneck**
  - Long user histories increase storage, search cost, and noise
  - Lightweight retrievers, memory compression, and candidate pruning are needed

- **Problem 2: Generator bottleneck**
  - Per-user fine-tuning is attractive but quickly becomes prohibitive
  - Parameter-efficient tuning and shared adapter designs are key

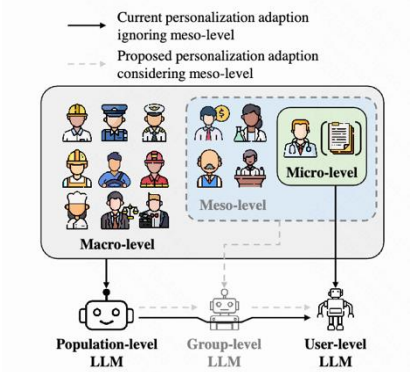
- **Problem 3: Memory bottleneck**
  - Search systems must decide what to keep, update, forget, and trust
  - Poor memory management causes drift, latency, and privacy risk

Challenges 3: Scalability

- **Representative works:**
  - **MTA & PROPER**
    - **Motivation**: Address the scalability challenges of personalized LLMs by leveraging model merging and parameter-efficient adaptation
    - **Contribution**: MTA proposes a merge-then-adapt framework for personalized LLMs, while PROPER serves as a representative baseline for personalized LoRA-based adaptation
    - **Method**: Train user-specific LoRA adapters locally, merge personalized knowledge across users, and adapt the merged model while preserving privacy-sensitive information



The overall framework of MTA



The overall framework of PROPER

[1] Li, Xiaopeng, et al. "MTA: A Merge-then-Adapt Framework for Personalized Large Language Model." arXiv preprint arXiv:2511.20072 (2025).
[2] Zhang, Linhai, et al. "Proper: A progressive learning framework for personalized large language models with group-level adaptation." Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025.

Challenges 3: Scalability

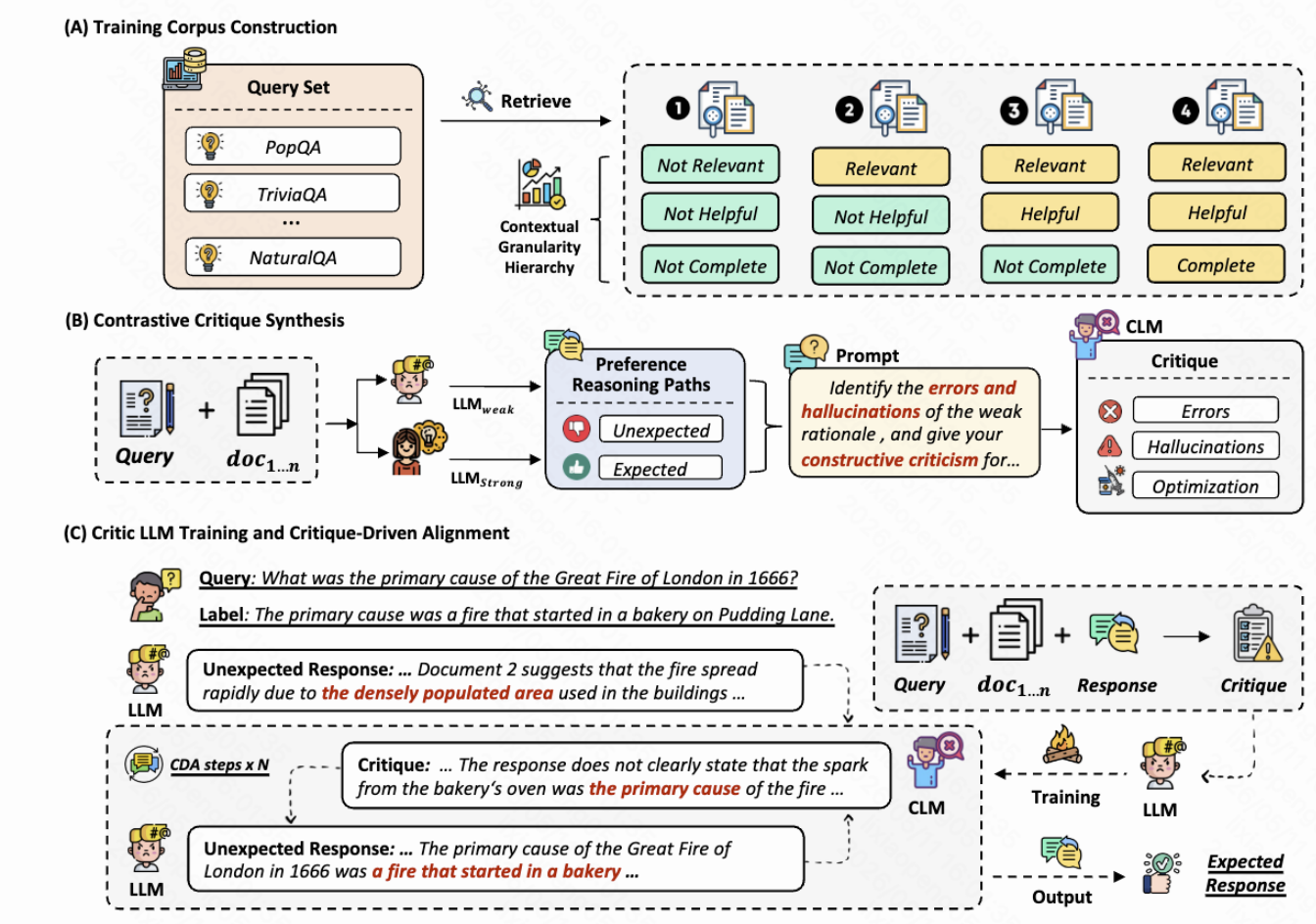
- **Representative works:**
  - **MTA &PROPER**
- **Potential future research area:**
  - **Multi-user LoRA or adapter frameworks**
    - Proposed more scalable framework for multiple users.
  - **Continual adaptation**
    - Update your preferences online, but avoid catastrophic forgetting and privacy breaches

Challenges 4: Ethics & Fairness

- **Problem 1: Bias amplification**
  - If profiles contain protected or correlated attributes, personalization may reinforce stereotypes or selectively expose groups to different content quality
- **Problem 2: User agency**
  - Users should be able to inspect, correct, delete, or constrain the profile data used for personalization
- **Problem 3: Fairness audits**
  - Evaluate whether personalization causes disparate outcomes, content bubbles, unfair recommendations, or unsafe inferences

Challenges 4: Ethics & Fairness

- **Representative works:**
  - **AlignRAG**
    - **Motivation**: Personalized systems may retrieve the right facts but generate answers that ignore or contradict them
    - **Contribution**: Introduce iterative critique or feedback to enforce alignment with retrieved content
    - **Method**: Compare reasoning trajectory against evidence, then correct reasoning drift before final generation



[1] Wei, Jiaqi, et al. "Retrieval is Not Enough: Enhancing RAG Reasoning through Test-Time Critique and Optimization." arXiv preprint arXiv:2504.14858 (2025).

Challenges 4: Ethics & Fairness

- **Representative works:**
  - **AlignRAG**

- **Potential future research area:**
  - **Fairness auditing for personalized systems**
    - Integrate fairness audits into personalized intelligence systems to detect whether personalization amplifies bias or produces discriminatory outputs
  - **Bias mitigation in personalized generation and recommendation**
    - Debias personalized recommendations and use adversarial training to remove signals from protected attributes such as gender or ethnicity

[1] Wei, Jiaqi, et al. "Retrieval is Not Enough: Enhancing RAG Reasoning through Test-Time Critique and Optimization." arXiv preprint arXiv:2504.14858 (2025).

Challenges 5: Multimodality

- **Problem:**
  - A text profile cannot fully encode style boards, reference photos, voice preferences, brand constraints, or interface behavior
  - Future systems need patch-, segment-, shot-, track-, and event-level retrieval, not only document-level retrieval
- **Representative Works**
  - ImageRAG: query-conditioned image retrieval -> visual conditioning -> grounded generation or reasoning
  - Multi-RAG: multimodal indexing -> modality-aware retrieval -> cross-modal fusion -> answer or action
  - AppAgent: screen understanding -> action selection -> tap/swipe/text execution
- **Potential future research area:**
  - Personalized multimodal-aware agents
  - Personalized hybrid multimodal retrieval

[1] Shalev-Arkushin, Rotem, et al. "Imagerag: Dynamic image retrieval for reference-guided image generation." arXiv preprint arXiv:2502.09411 (2025).
[2] Mao, Mingyang, et al. "Multi-rag: A multimodal retrieval-augmented generation system for adaptive video understanding." arXiv preprint arXiv:2505.23990 (2025).
[3] Zhang, Chi, et al. "Appagent: Multimodal agents as smartphone users." Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. 2025.

Challenges 6: Agent Planning

- **Problem:**
  - Current LLM agents mostly plan in a one-size-fits-all manner
  - Personalization is rarely integrated into the planning logic that decides next actions
  - Existing personalized agents are mostly isolated domain-specific cases, not general agentic benchmarks
- **Representative Works**
  - CHAs: user health profile -> personalized healthcare assistance and diagnosis
  - ExpertAgent: learner profile -> personalized learning content and strategies
  - RecAgent: hierarchical user memory -> cross-domain personalized recommendation planning
- **Potential future research area:**
  - Personalized long-term memory for agent planning
  - Dynamically evolving user-aligned decision systems
  - Meta-learning for user-specific planning strategies

[1] Abbasian, Mahyar, et al. "Conversational health agents: A personalized llm-powered agent framework." arXiv preprint arXiv:2310.02374 (2023).
[2] Zhu, Binrong, Guiran Liu, and Nina Jiang. "ExpertAgent: Enhancing Personalized Education through Dynamic Planning and Retrieval-Augmented Long-Chain Reasoning." arXiv preprint arXiv:2510.07456 (2025).
[3] Wang, Lei, et al. "User behavior simulation with large language model-based agents." ACM Transactions on Information Systems 43.2 (2025): 1-37.

Takeaways

➢Research agenda for the next few years

01 Holistic benchmarks
Evaluate pre-retrieval, retrieval, generation, memory, and planning together.

02 Privacy-preserving personalization
Build edge-cloud, federated, encrypted, and auditable personalization stacks.

03 Multi-user efficient adaptation
Develop shared adapter systems that scale without erasing individual differences.

04 Fair and coherent systems
Audit bias, expose user controls, and align reasoning with retrieved evidence.

05 Multimodal retrieval granularity
Retrieve patches, segments, shots, tracks, and GUI traces, not only documents.

06 Personalized planning
Integrate user models into agent policies, tool selection, and long-horizon decisions.

Agenda

1. Introduction
9:00 – 9:50



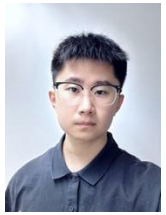
Pengyue Jia

2. Pre-retrieval
9:50 – 10:30



Pengyue Jia

3. Retrieval
11:00 – 11:50



Xiaopeng Li

4. Generation
11:50 – 12:30



Wenlin Zhang

5. From RAG to Agent
14:00 – 15:00



Derong Xu

6. Evaluation
15:00 – 15:30



Ziwei Liu

7. Future & Conclusion
16:00 – 16:45



Ziwei Liu

8. Hands on Projects
16:45 – 17:30



Pengyue Jia

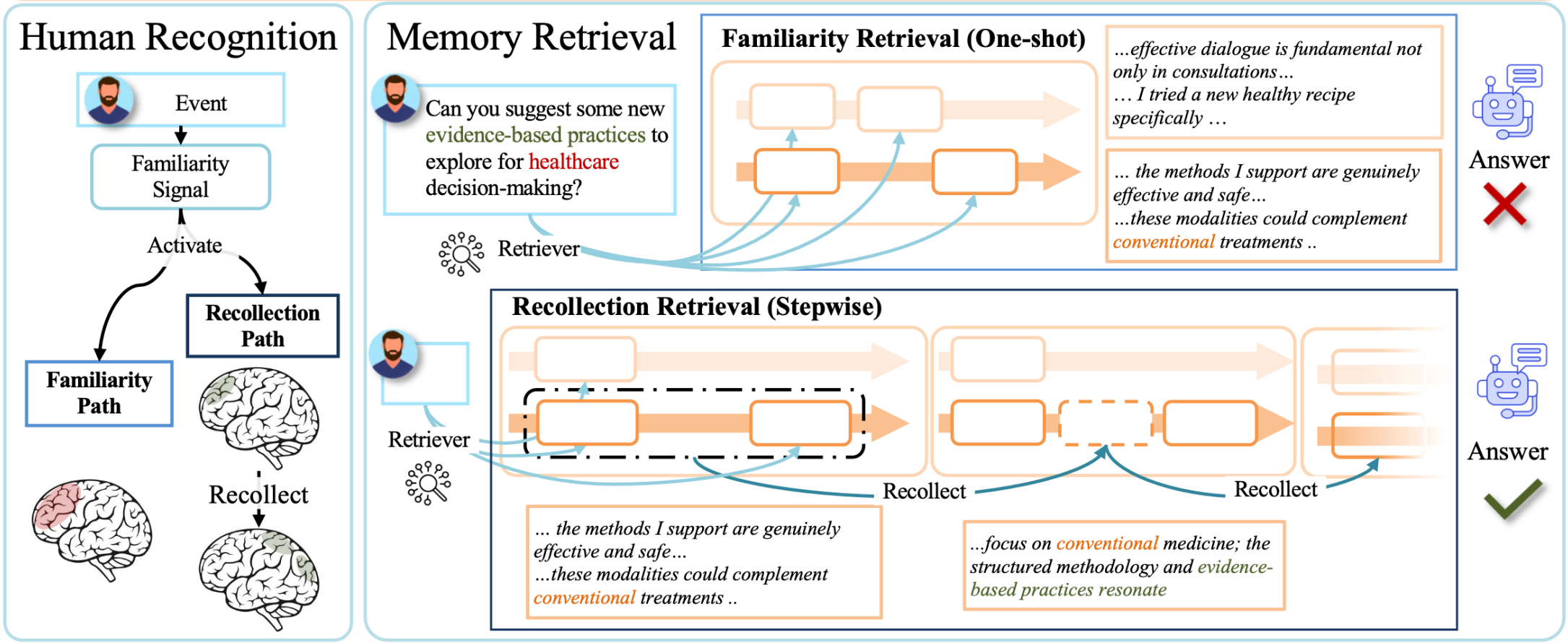
Hands on Project (RF-Mem)

SIGIR2026 20-24 July MELBOURNE | NAARM AUSTRALIA

EVOKING USER MEMORY: PERSONALIZING LLM VIA RECOLLECTION-FAMILIARITY ADAPTIVE RETRIEVAL

Yingyi Zhang[1,2,*], **Junyi Li**[2,*], **Wenlin Zhang**[2], **Penyue Jia**[2], **Xianneng Li**[1,†], **Yichao Wang**[3,†], **Derong Xu**[2,4], **Yi Wen**[2], **Huifeng Guo**[3], **Yong Liu**[3], **Xiangyu Zhao**[2,†]
[1]Dalian University of Technology, [2]City University of Hong Kong,
[3]Huawei Technologies Ltd., [4]University of Science and Technology of China
xianneng@dlut.edu.cn,wangyichao5@huawei.com,xianzhao@cityu.edu.hk

Hands on Projects (RF-Mem)



Hands on Projects (RF-Mem)

Visit https://tinyurl.com/prag-sigir26 or Scan the QR Code to get the jupyter notebook

RF-Mem: Recollection–Familiarity Adaptive Retrieval for Personalized LLMs

Open in Colab

SIGIR 2026 Tutorial — *Personalized RAG and Agents* · Hands-on Session

This notebook is a 45-minute interactive walk-through of:

> **Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval** *ICLR 2026* — arXiv:2603.09250 · GitHub

What you will do

1. **Understand the personalised long-term memory task** using LongMemEval.
2. **Build a vanilla dense-retrieval baseline** (the "Familiarity" path).
3. **Re-implement RF-Mem's Recollection path** — pyramid clustering with α-mix query expansion.
4. **Implement the uncertainty-aware router** that picks fast vs. slow retrieval per query.
5. **Compare** the three methods on Recall@k and latency, and inspect qualitative cases.

> ⚠️ This notebook is sized for the Colab free tier (T4GPU is enough). We use a small embedding model and a small evaluation subset so everything finishes in a few minutes.

Conclusion

- We summarized **personalization** techniques from RAG to advanced LLM-based agents, covering **pre-retrieval, retrieval, and generation** stages
- We reviewed recent **literature, datasets**, and **evaluation metrics**
- Key challenges include **effective evaluation, ethical/privacy concerns, and scalability, etc.**
- Future research should prioritize **Holistic benchmarks, privacy-preserving methods, lightweight frameworks, etc.**

CityU AML Lab

Tutorial Homepage

Survey Paper