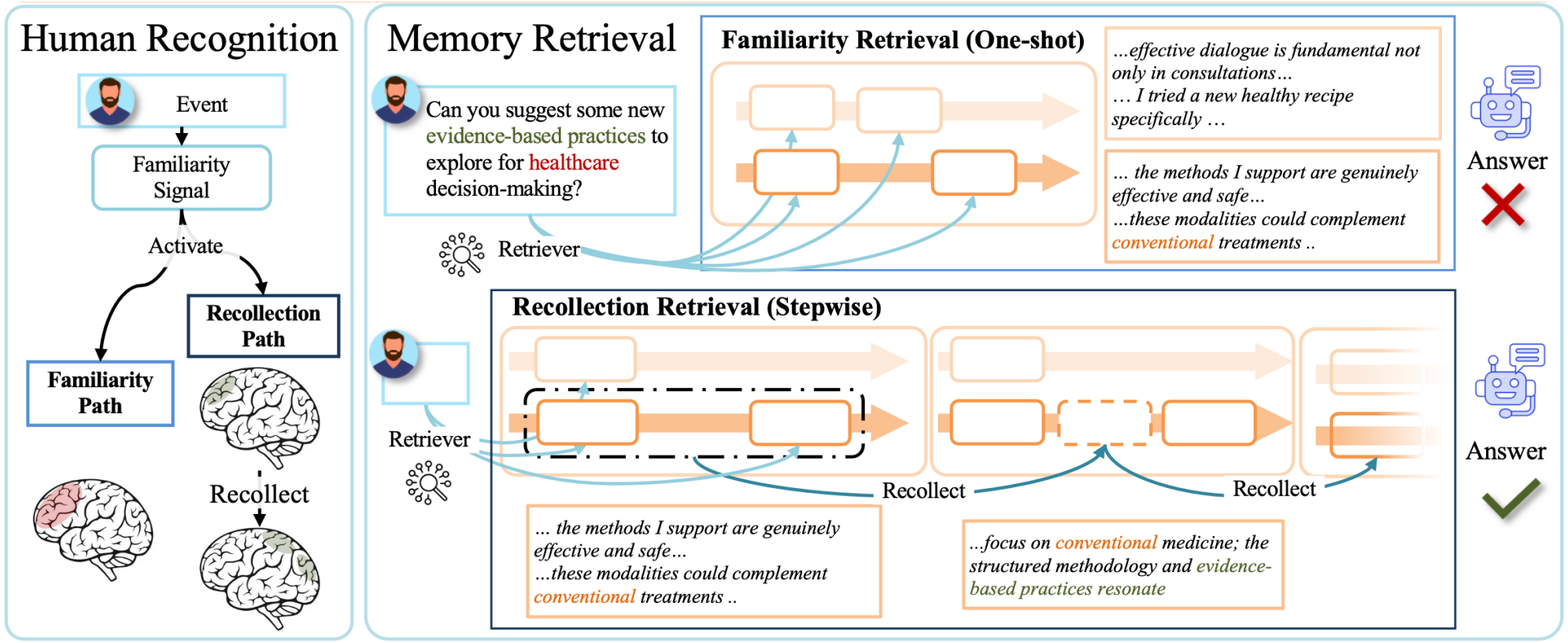
Evoking User Memory: Personalizing LLM via Recollection-Familiarity Adaptive Retrieval

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

Content



Motivation & Problem

LLM recently act as personal assistant to support people's life to answer personalized question

Event: Writing Email to apply for visa

How should I email them to follow up on my visa application?

Event: Develop personal webpage

Help me to update my personal website based on this paper.

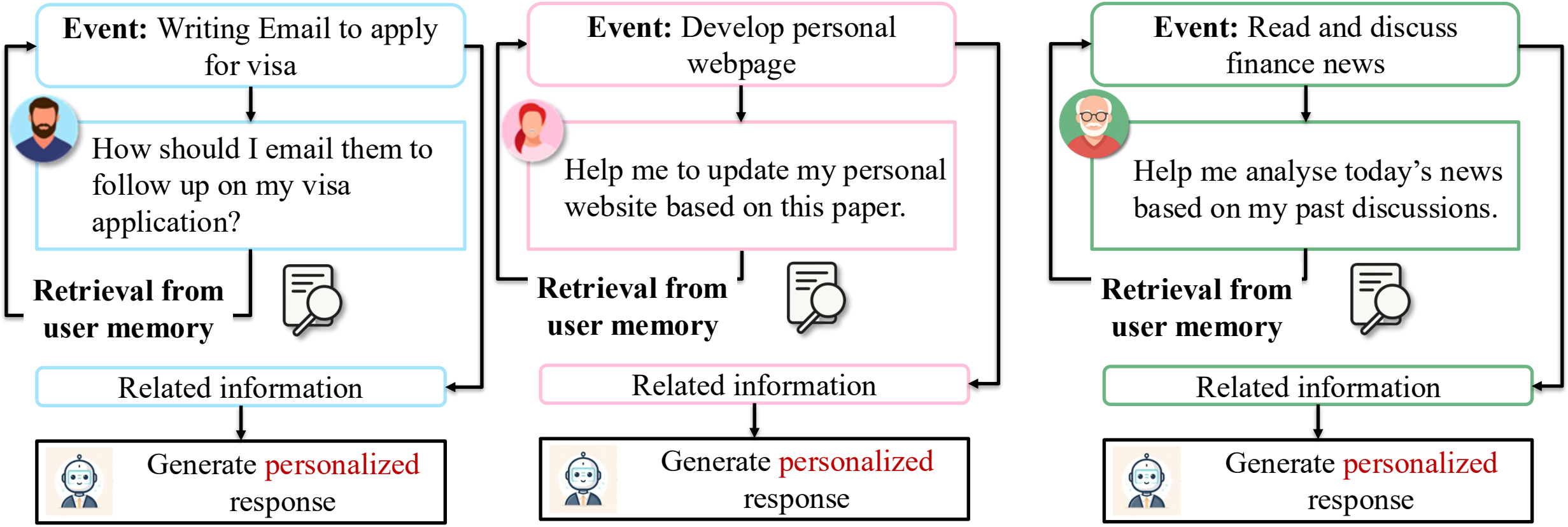
Event: Read and discuss finance news

Help me analyse today's news based on my past discussions.

People Using LLM at Any Time!

Motivation & Problem

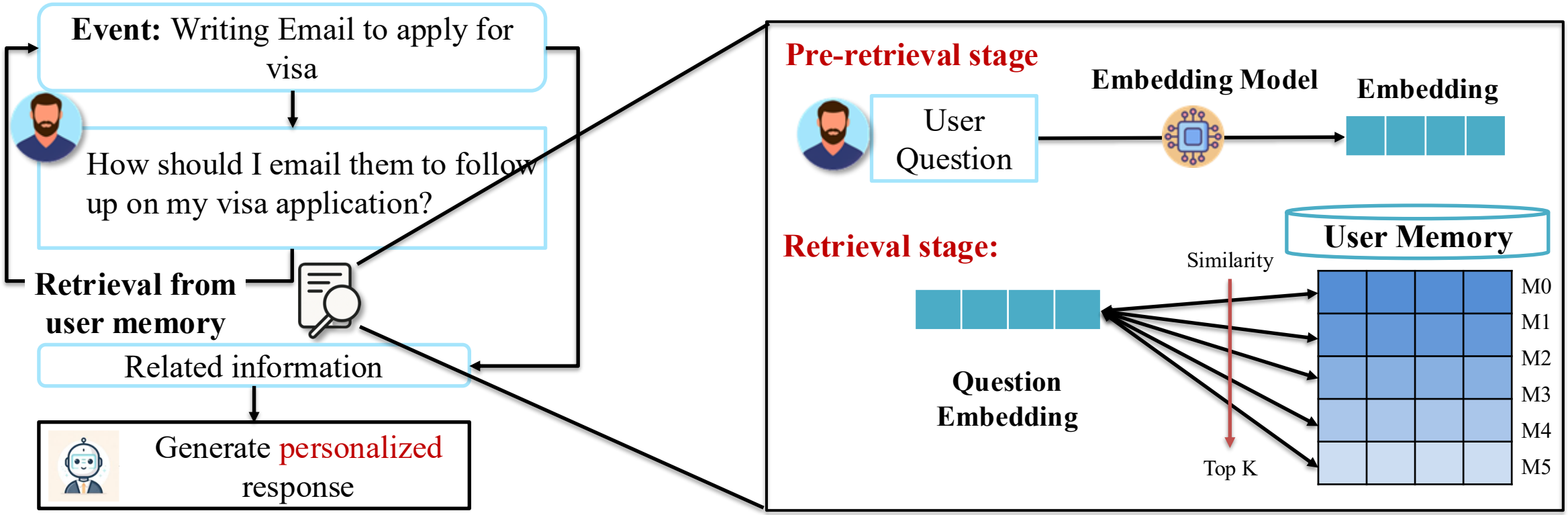
For answer these question, Retrieval-Augmented Generation (RAG) has emerged as a pivotal paradigm for generate personalized response



People Using LLM at Any Time!

Motivation & Problem

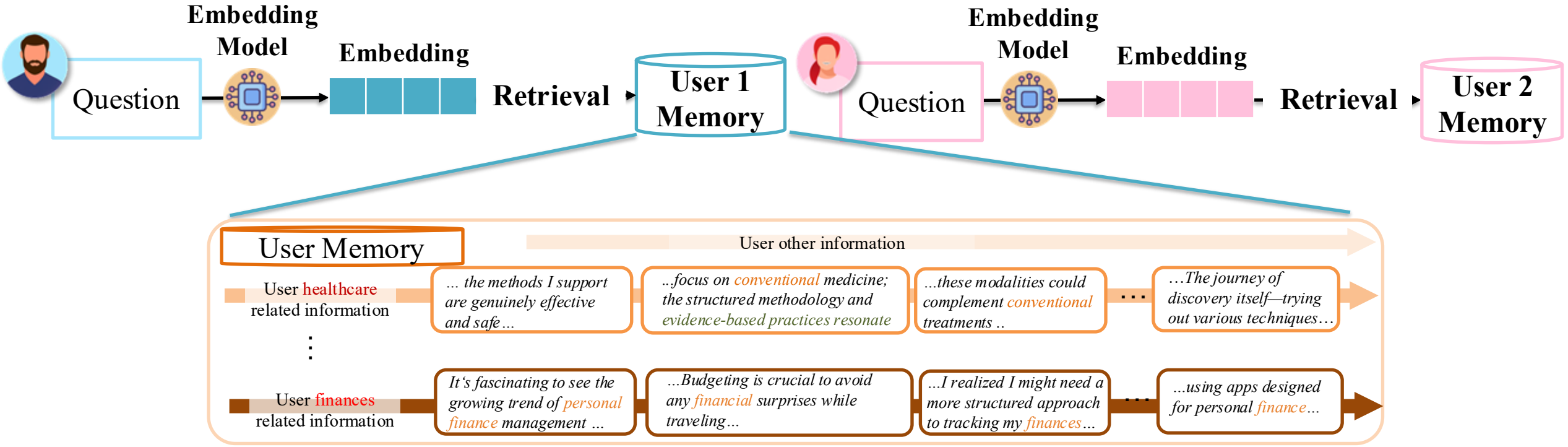
For answer these question, Retrieval-Augmented Generation (RAG) has emerged as a pivotal paradigm for generate personalized response



People Using LLM at Any Time!

Motivation & Problem

In personal copra, it have two features.

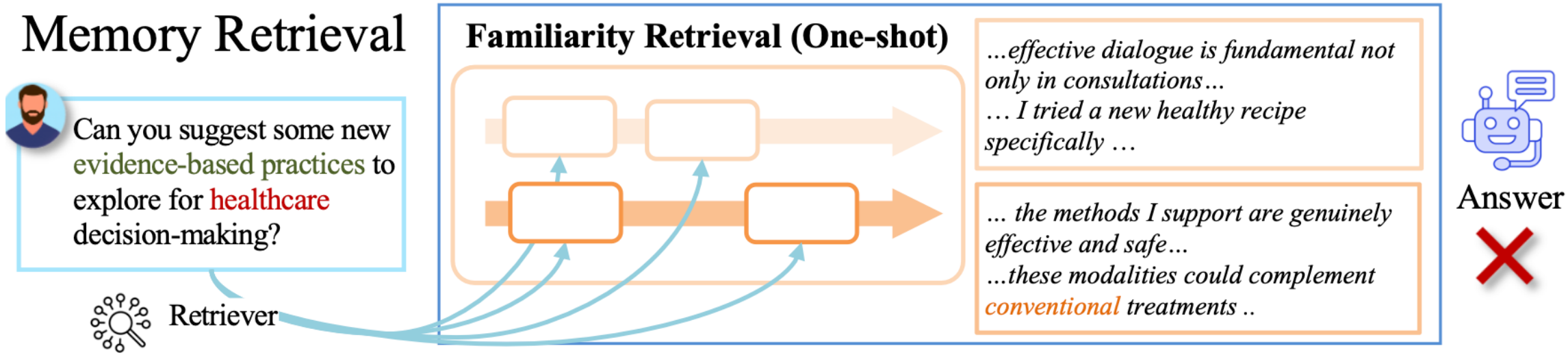


- **Over time, each user memory corpus spanning multiple topics.**
- **Only a small portion of memories are relevant to the current query.**

Motivation & Problem

Motivation

- Existing retrieval methods mainly **rely on one-shot similarity search**, which captures only surface-level matches.

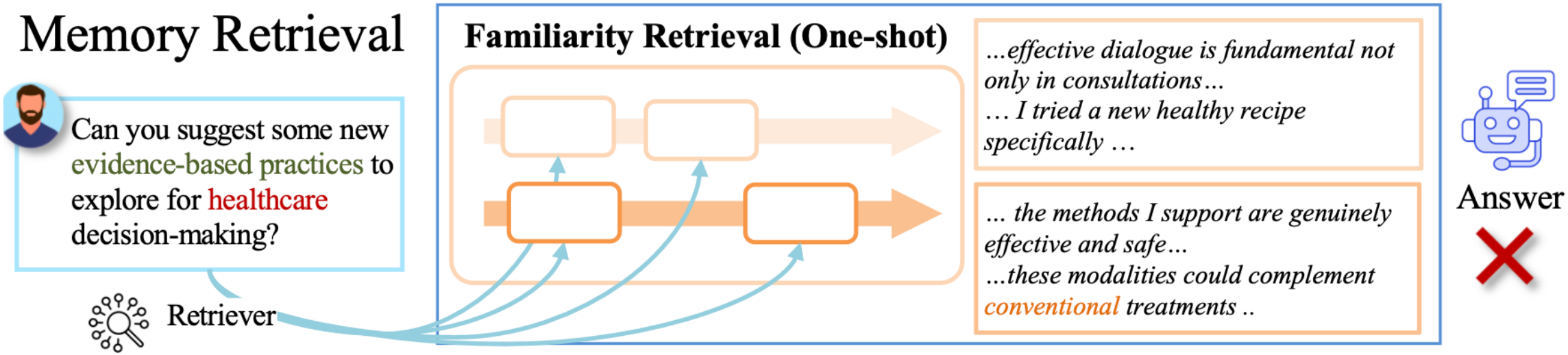


- **Complex personalized queries often require recovering contextual evidence scattered across user memories.**

Motivation & Problem

Motivation

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

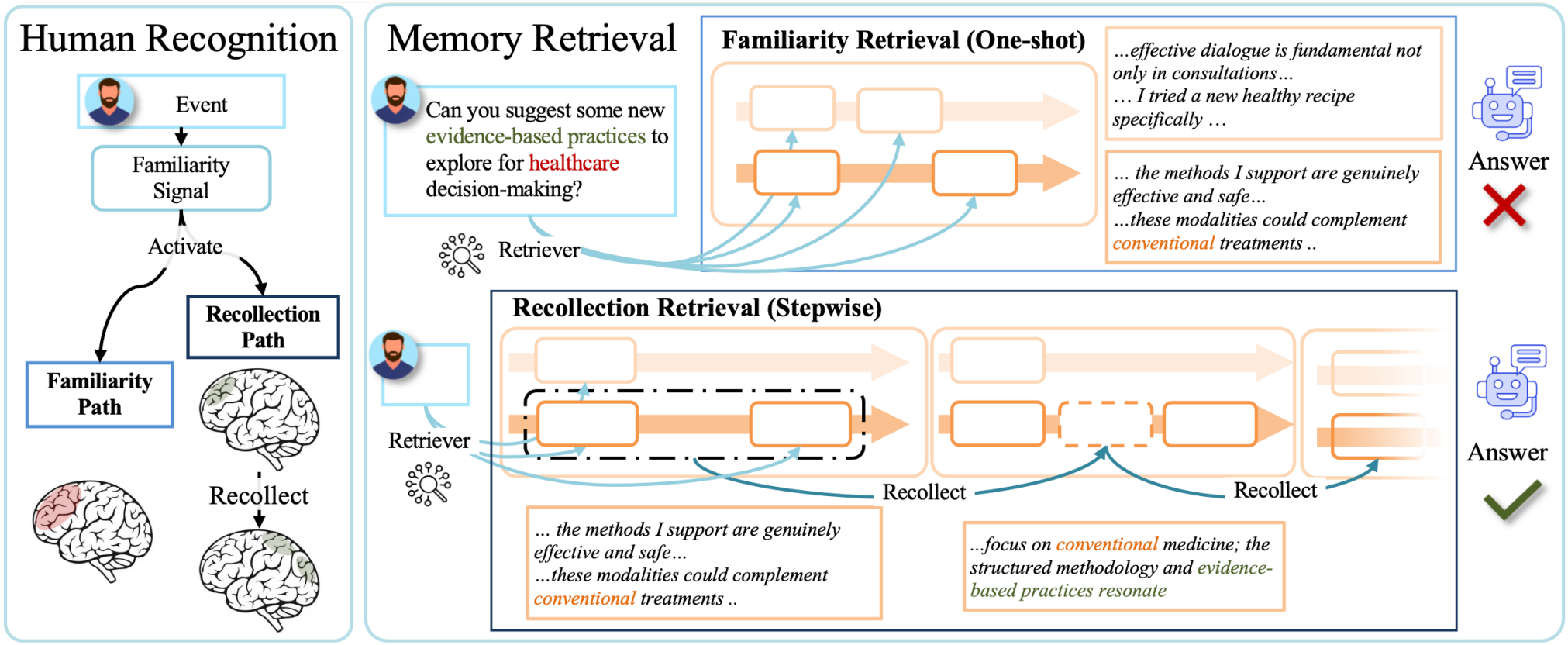


- **Current systems lack both the ability to perform recollection retrieval and mechanisms to adaptively switch between the dual retrieval paths, leading to either insufficient recall or the inclusion of noise.**

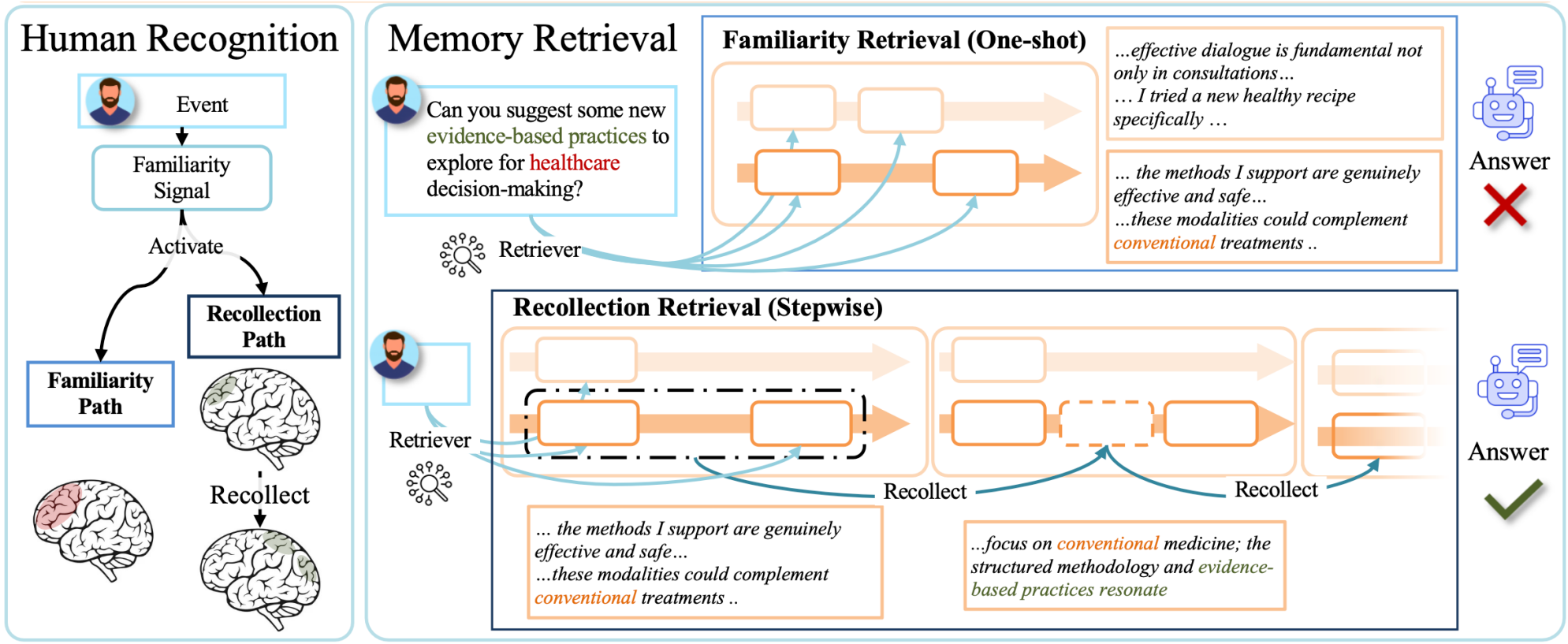
Motivation & Problem

Therefore

- **we propose RF-Mem (Recollection–Familiarity Memory Retrieval), a familiarity uncertainty-guided dual-path memory retriever.**



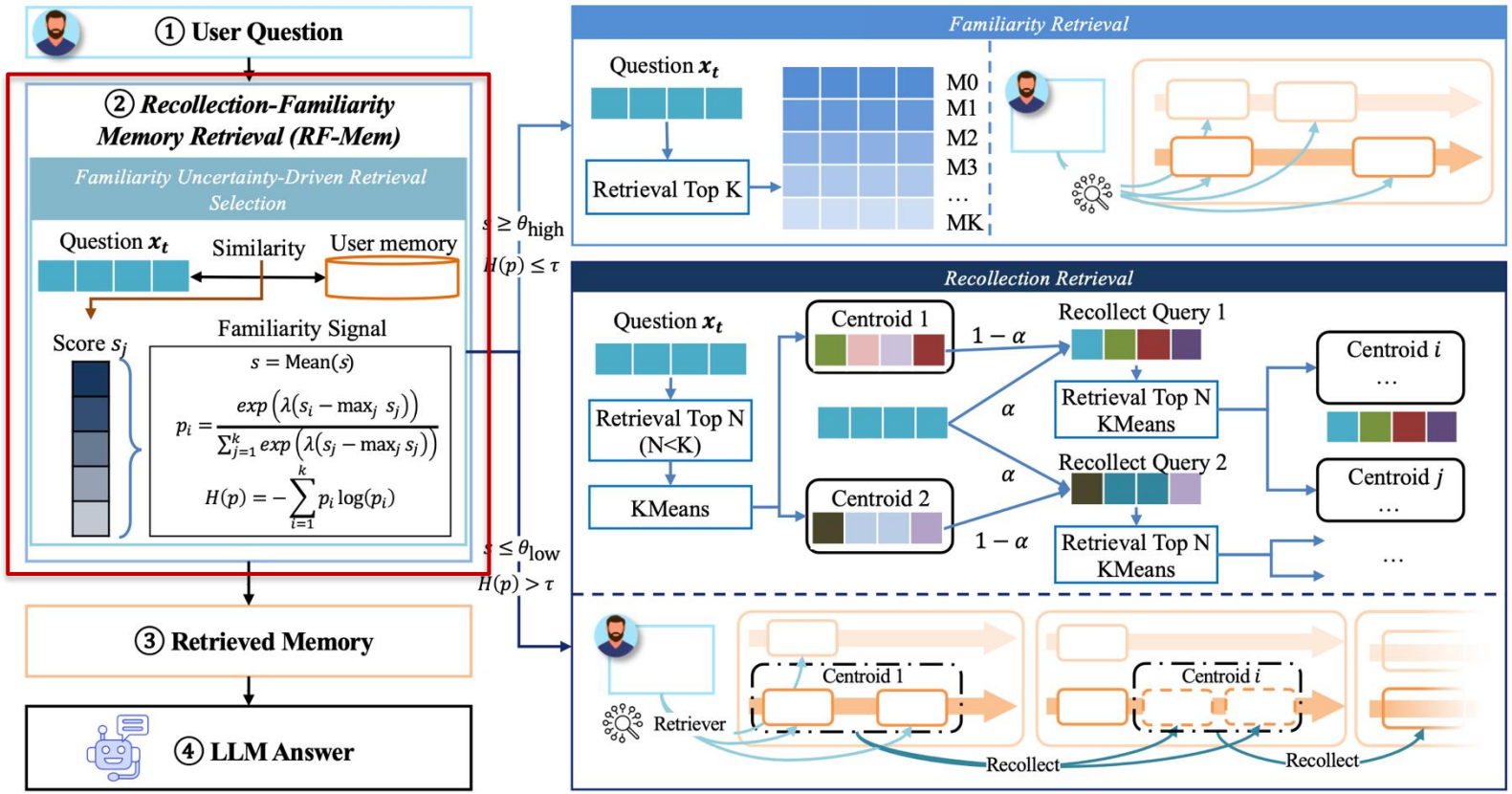
Content



RF-Mem: Overview

RF-Mem Framework

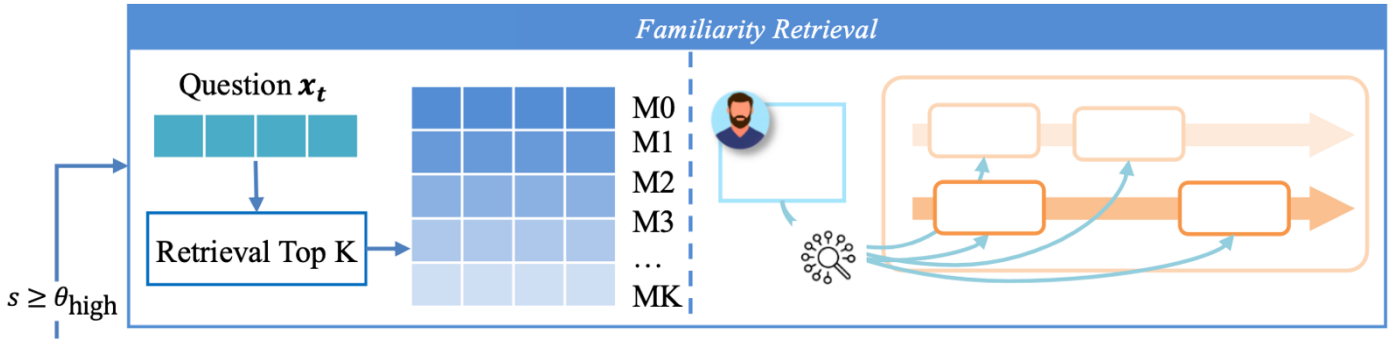
- RF-Mem **first performs a probe retrieval to estimate a familiarity signal** using the mean similarity score and entropy of retrieved candidates.



RF-Mem: Overview

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.

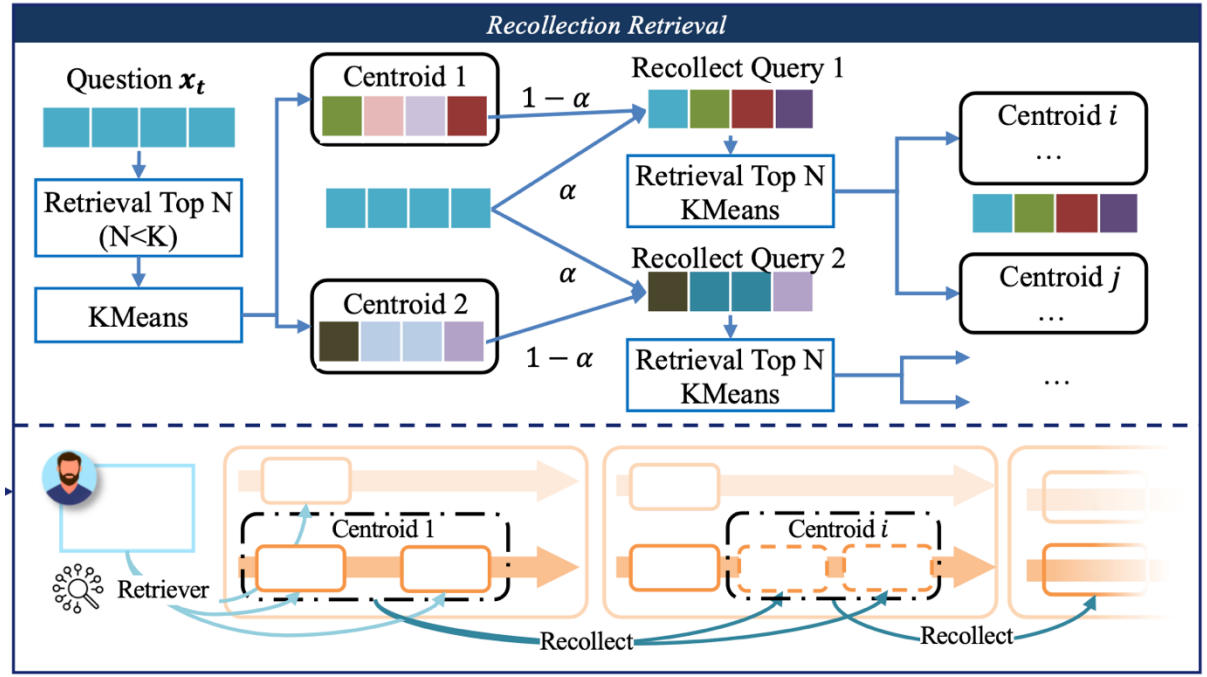


RF-Mem: Overview

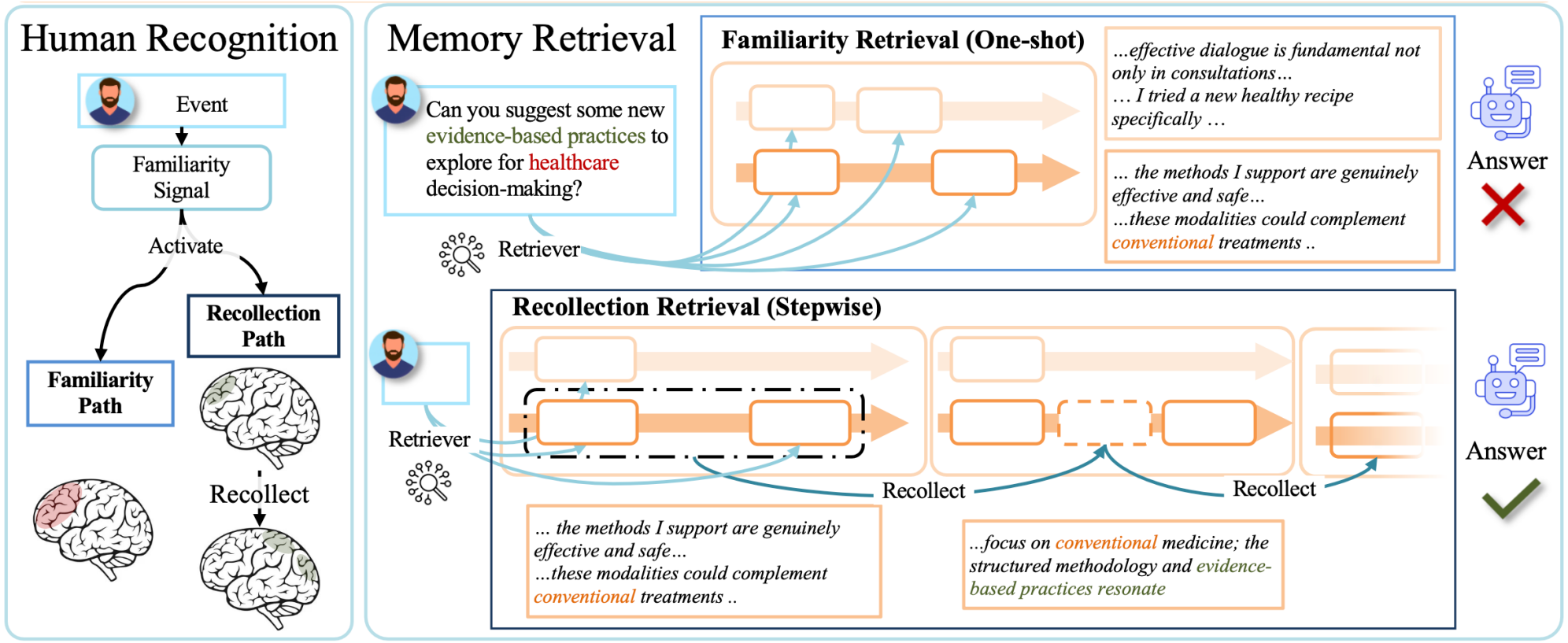
RF-Mem Framework

- 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



Content



Experiment Settings

Dataset:

- PersonaMem[1]: includes multiple simulated user-LLM interaction histories over 7 real-world tasks
- LongMemEval[2]: targets long-term interactive memory, with two subsets: -s and –m
- PersonaBench[3]: simulates user-specific queries over synthetic private data

Evaluation metrics:

- Accuracy
- Recall@K
- NDCG@K

Table 7: Dataset Statistics across Different Memory Corpora.

Dataset name	32k memory corpus	128k memory corpus	1M memory corpus
# of samples	589	2727	2674
Avg tokens of question	464.6	416.3	415.1
Avg tokens of memory	24193.2	15185.1	911733.4
# of Revisit Reasons	99	269	235
# of Track Evolution	139	341	225
# of Latest Prefs	17	866	768
# of Aligned Recs	55	349	280
# of New Scenarios	57	213	295
# of Shared Facts	129	171	144
# of New Ideas	93	518	727

Table 9: Statistics of the PersonaBench dataset across users.

User Id	# of Queries	# of Corpus	# of Conversations with friends	# of User-AI Conversation	# of User e-comerce purchase histories
1	48	110	84	23	3
2	43	90	78	8	4
3	42	64	51	12	1
4	46	85	71	14	0
5	44	84	59	21	4
6	40	94	79	14	1
Sum	263	527	422	92	13
Avg	**43.83**	**87.83**	**70.33**	**15.33**	**2.17**

Table 10: Statistics of the LongMemEval dataset, with different sizes of associated memory corpora.

Statistic	LongMemEval-s	LongMemEval-m
Total Questions	500	500
Total Session-level Memory	25,112	250,948
Min Session-level Memory per Question	39	501
Max Session-level Memory per Question	66	506
Avg Session-level Memory per Question	**50.22**	**501.90**

[1] Jiang, Bowen, et al. "Know Me, Respond to Me: Benchmarking LLMs for Dynamic User Profiling and Personalized Responses at Scale." Second Conference on Language Modeling.
[2] Wu, Di, et al. "LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory." The Thirteenth International Conference on Learning Representations. ICLR 2025
[3] Tan, Juntao, et al. "Personabench: Evaluating ai models on understanding personal information through accessing (synthetic) private user data." Findings of the Association for Computational Linguistics: ACL 2025

Overall Performance

Method	Retri Time	Avg. Tokens	Revisit Reasons	Track Evolution	Latest Prefs	Aligned Recs	New Sce-narios	Shared Facts	New Ideas	Overall
32K memory corpus data										
Zero Memory	NA	464.6	0.7273	0.6259	0.1765	0.2182	0.2105	0.2326	0.1183	0.3854
Full Context	NA	24657.8	0.9394	**0.7194**	**0.7647**	0.7455	0.5614	0.5039	0.1828	0.6129
Dense Retrieval	3.14ms	3515.9	0.9091	0.6475	0.6471	0.6364	0.5614	0.5426	0.2151	0.5908
Recol. (ours)	7.09ms	3711.1	**0.9495**	0.6547	0.7059	**0.7818**	0.5965	0.5194	**0.2688**	0.6214
RF-Mem (ours)	5.09ms	3566.6	**0.9495**	0.6619	0.7059	**0.7818**	**0.6140**	**0.5659**	**0.2688**	**0.6350***
128K memory corpus data										
Zero Memory	NA	416.3	0.6766	0.6422	0.2136	0.2751	0.1925	0.2281	0.1737	0.3124
Full Context	NA	115601.4	0.5613	0.3930	0.2783	0.3868	0.2770	0.3977	0.1795	0.3231
Dense Retrieval	3.24ms	3540.1	0.7881	0.6804	0.5346	0.5330	0.3662	0.6082	0.3069	0.5259
Recol. (ours)	7.86ms	3680.3	**0.8141**	0.6716	0.5254	0.5301	0.3765	0.6140	**0.3263**	0.5288
RF-Mem (ours)	4.27ms	3565.5	0.8030	**0.6862**	**0.5427**	**0.5358**	**0.4131**	**0.6257**	**0.3263**	**0.5394***
1M memory corpus data										
Zero Memory	NA	415.1	0.6000	0.6178	0.1797	0.3179	0.1831	0.2569	0.1816	0.2730
Full Context	NA	912148.5	OOC	OOC	OOC	OOC	OOC	OOC	OOC	OOC
Dense Retrieval	4.42ms	3816.1	0.7702	**0.6933**	**0.4544**	0.4464	0.3085	0.5903	0.3040	0.4518
Recol. (ours)	8.12ms	3847.4	0.7532	0.6800	0.4440	0.4500	**0.3593**	0.5833	0.3136	0.4544
RF-Mem (ours)	6.28ms	3827.8	**0.7787**	0.6889	0.4492	**0.4536**	0.3390	**0.6111**	**0.3150**	**0.4589***

1. **RF-Mem delivers the best overall accuracy at every corpus scale while keeping inputs compact.**

Overall Performance

Metrics	Recall@5						Recall@10					
Method	Time	Basic Info	Social Info	Pref Easy	Pref Hard	Overall	Time	Basic Info	Social Info	Pref Easy	Pref Hard	Overall
multi-qa-MiniLM-L6-cos-v1												
Famili.	8.40ms	0.4515	0.4852	0.4904	0.3659	0.4484	13.68ms	0.5879	0.6220	**0.6442**	0.5561	0.5964
Recol.	9.65ms	0.4379	0.4903	**0.5128**	**0.3854**	0.4491	17.29ms	**0.5924**	**0.6859**	0.5659	**0.6267**	0.6062
RF-Mem	9.16ms	**0.4788**	**0.5091**	0.4872	**0.3854**	**0.4701**	15.22ms	**0.5924**	0.6799	0.5707	**0.6267**	**0.6071**
all-mpnet-base-v2												
Famili.	7.64ms	0.4242	0.2730	0.4487	**0.4049**	0.3887	10.55ms	0.5409	**0.4434**	**0.6795**	0.5366	0.5333
Recol.	10.94ms	0.4333	**0.2918**	**0.4583**	0.4000	0.3976	13.23ms	**0.6000**	0.4365	0.6378	0.5220	0.5527
RF-Mem	8.33ms	**0.4515**	0.2730	0.4487	0.4000	**0.4009**	10.55ms	0.5955	0.4384	0.6378	**0.5463**	**0.5553**
BAAI/bge-base-en-v1.5												
Famili.	8.92ms	0.3970	0.3204	**0.4583**	**0.3268**	0.3738	10.19ms	0.5121	**0.4748**	**0.5673**	**0.4585**	0.5002
Recol.	12.14ms	0.3833	**0.3619**	0.4327	0.3171	0.3722	20.71ms	0.5212	**0.4748**	**0.5673**	**0.4585**	0.5046
RF-Mem	10.14ms	**0.4015**	**0.3619**	0.4487	0.3220	**0.3836**	18.13ms	**0.5303**	**0.4748**	**0.5673**	**0.4585**	**0.5089**

Method	LongMemEval-S				LongMemEval-M			
	Recall@5	Recall@10	Recall@50	Time	Recall@5	Recall@10	Recall@50	Time
multi-qa-MiniLM-L6-cos-v1								
Fami.	0.7136	0.8282	0.9761	24.91ms	0.4177	0.5465	0.7518	27.72ms
Recol.	0.7351	0.8425	**1.0000**	50.62ms	0.4368	0.5585	0.7590	57.93ms
RF-Mem	**0.7375**	**0.8473**	**1.0000**	39.58ms	**0.4391**	**0.5609**	**0.7613**	41.22ms
all-mpnet-base-v2								
Fami.	0.7303	0.8353	0.9832	27.25ms	0.4176	0.5489	0.7637	33.18ms
Recol.	**0.7398**	0.8305	**0.9952**	51.79ms	0.4386	0.5871	0.7422	62.11ms
RF-Mem	**0.7398**	**0.8377**	**0.9952**	42.39ms	**0.4391**	**0.5894**	**0.7684**	50.80ms
BAAI/bge-base-en-v1.5								
Fami.	0.7924	0.8926	**1.0000**	29.65ms	0.4964	0.6611	0.8305	30.77ms
Recol.	0.8162	0.9165	**1.0000**	43.65ms	0.5131	**0.6635**	0.8234	58.05ms
RF-Mem	**0.8186**	**0.9189**	**1.0000**	37.34ms	**0.5155**	**0.6635**	**0.8329**	44.74ms

1. **RF-Mem delivers the best overall accuracy at every corpus scale while keeping inputs compact.**
2. **RF-Mem achieves the most balanced and robust performance across retrievers in memory retrieval.**

Adaptive Study

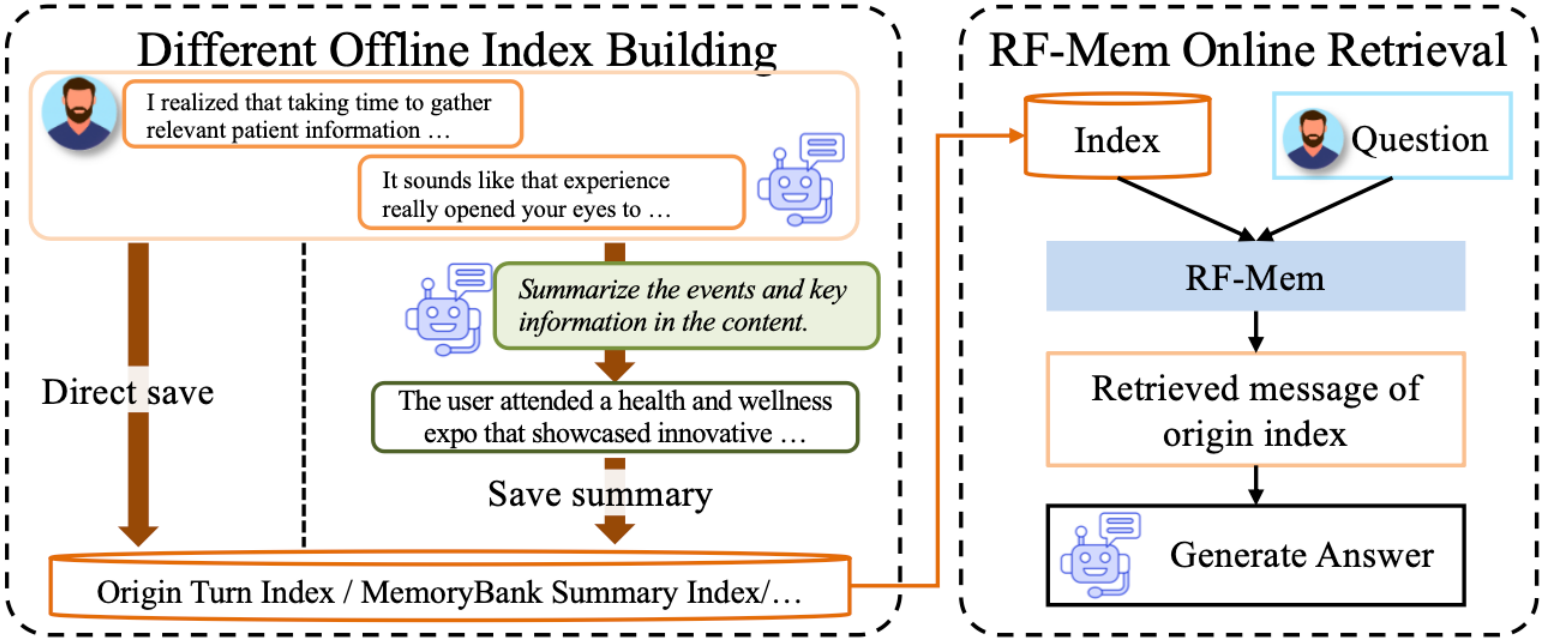


Table 4: Results by using MemoryBank summary index on PersonaMem (32K corpus).

Method	Avg. Tokens	Revisit Reasons	Track Evolution	Latest Prefs	Aligned Recs	New Scenarios	Shared Facts	New Ideas	Overall
MemoryBank summary index									
Familiarity	1267.6	0.7475	0.6187	0.5882	0.5818	0.3333	0.4341	0.1505	0.4941
Recollection	1441.8	0.8182	0.6259	0.5294	0.6909	0.4737	0.4031	0.1398	0.5212
RF-Mem	1421.8	0.8384	0.6259	0.5294	0.6545	0.4211	0.4419	0.1828	**0.5314**

1. **RF-Mem is modular and can be layered on top of heterogeneous memory indices**

Adaptive Study

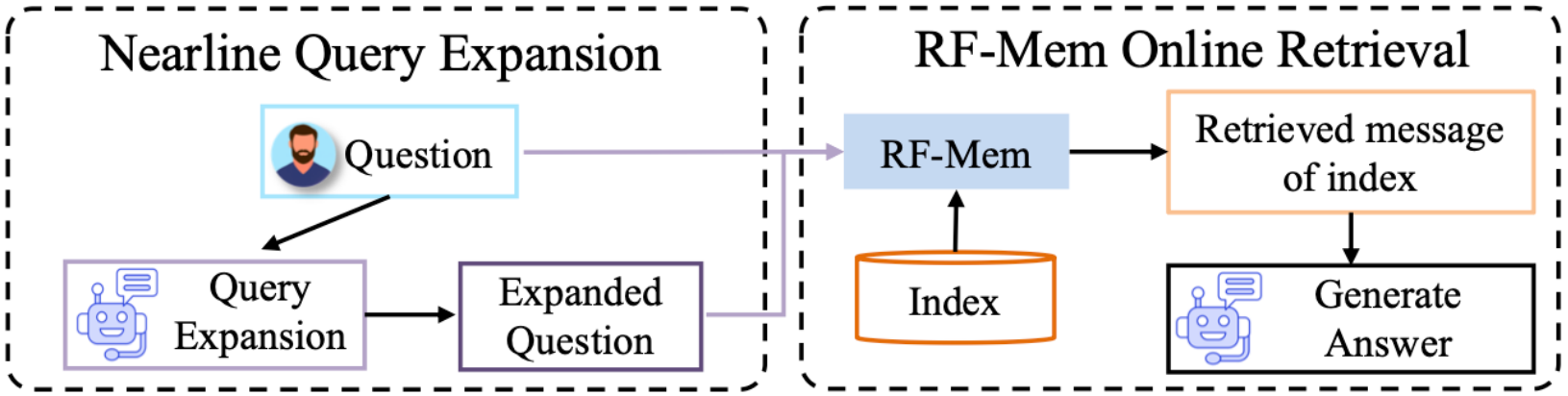


Table 5: Results by using HyDE query expansion method on PersonaBench.

Metrics	Recall@5					Recall@10				
HyDE	Basic Info	Social Info	Pref Easy	Pref Hard	Overall	Basic Info	Social Info	Pref Easy	Pref Hard	Overall
multi-qa-MiniLM-L6-cos-v1										
Famili.	0.3106	0.3909	0.4615	0.3122	0.3464	0.5000	0.4991	0.5737	0.5220	0.5120
Recol.	0.3015	0.4135	0.4615	0.3171	0.3482	0.4909	0.5028	0.5929	0.4878	0.5046
RF-Mem	0.3061	0.4135	0.4615	0.3171	**0.3504**	0.5091	0.5028	0.5929	0.5220	**0.5194**

1. **RF-Mem is modular and can be layered on top of heterogeneous memory indices**
2. **RF-Mem is modular and can be seamlessly integrated into nearline expansion pipelines.**

Adaptive Study

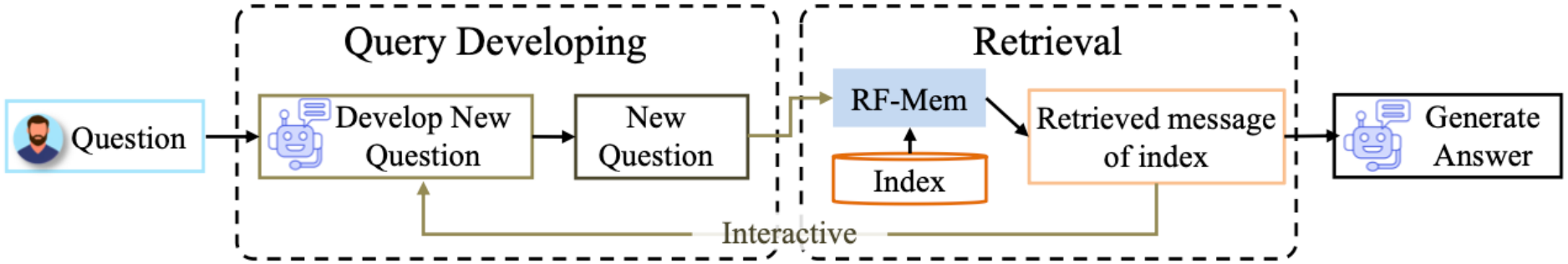
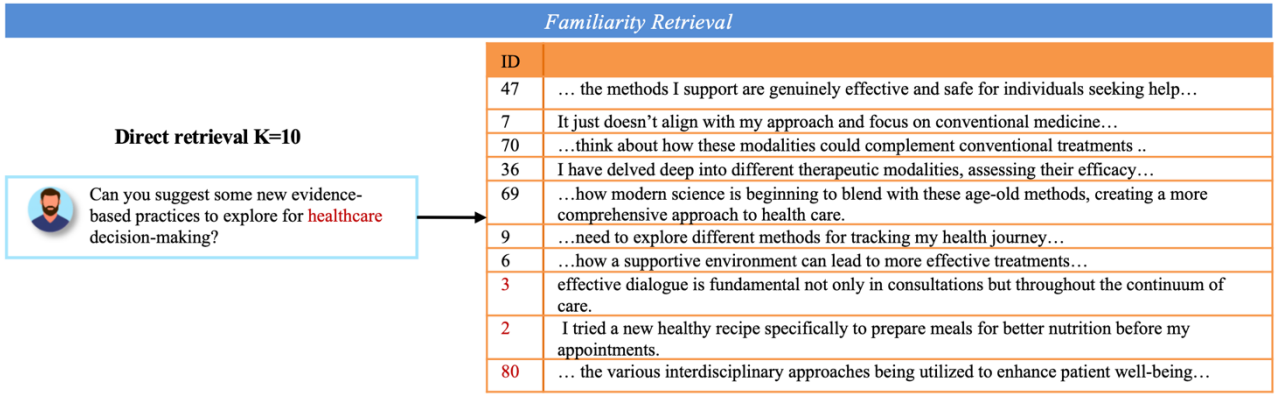


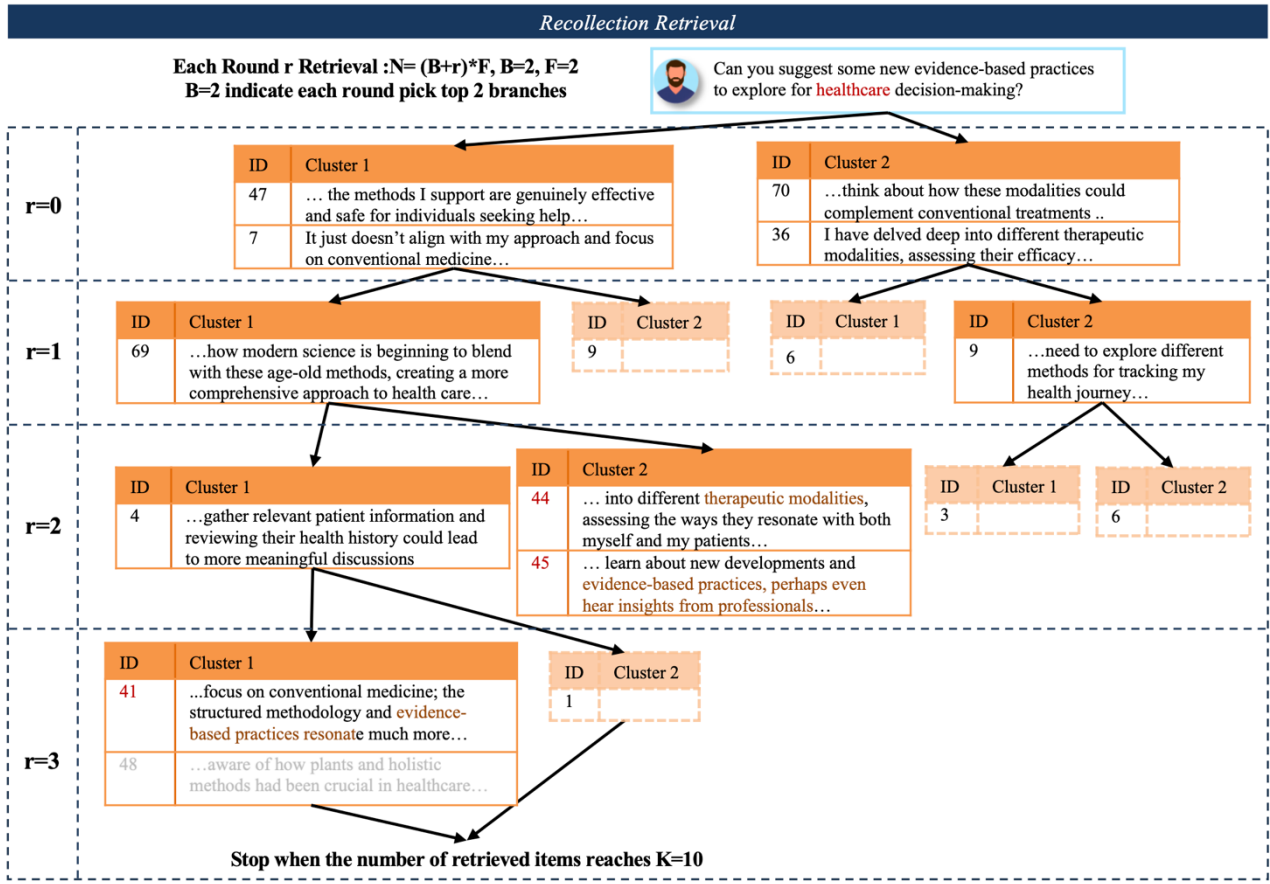
Table 6: Results by using Search-o1 iterative retrieval on PersonaMem (32K corpus).

Search-o1	Avg. Tokens	Revisit Reasons	Track Evolution	Latest Prefs	Aligned Recs	New Scenarios	Shared Facts	New Ideas	Overall
Search-o1									
Familiarity	4948.7	0.8687	0.6259	0.5882	0.6545	0.5614	0.5271	0.2581	0.5823
Recollection	5103.9	0.8990	0.6619	0.6471	0.7091	0.5789	0.4961	0.2796	0.6010
RF-MEM	5158.2	0.9293	0.6978	0.6471	0.7455	0.5789	0.5349	0.2043	**0.6146**

1. **RF-Mem is modular and can be layered on top of heterogeneous memory indices**
2. **RF-Mem is modular and can be seamlessly integrated into nearline expansion pipelines.**
3. **RF-Mem retains its effectiveness in iterative RAG settings.**

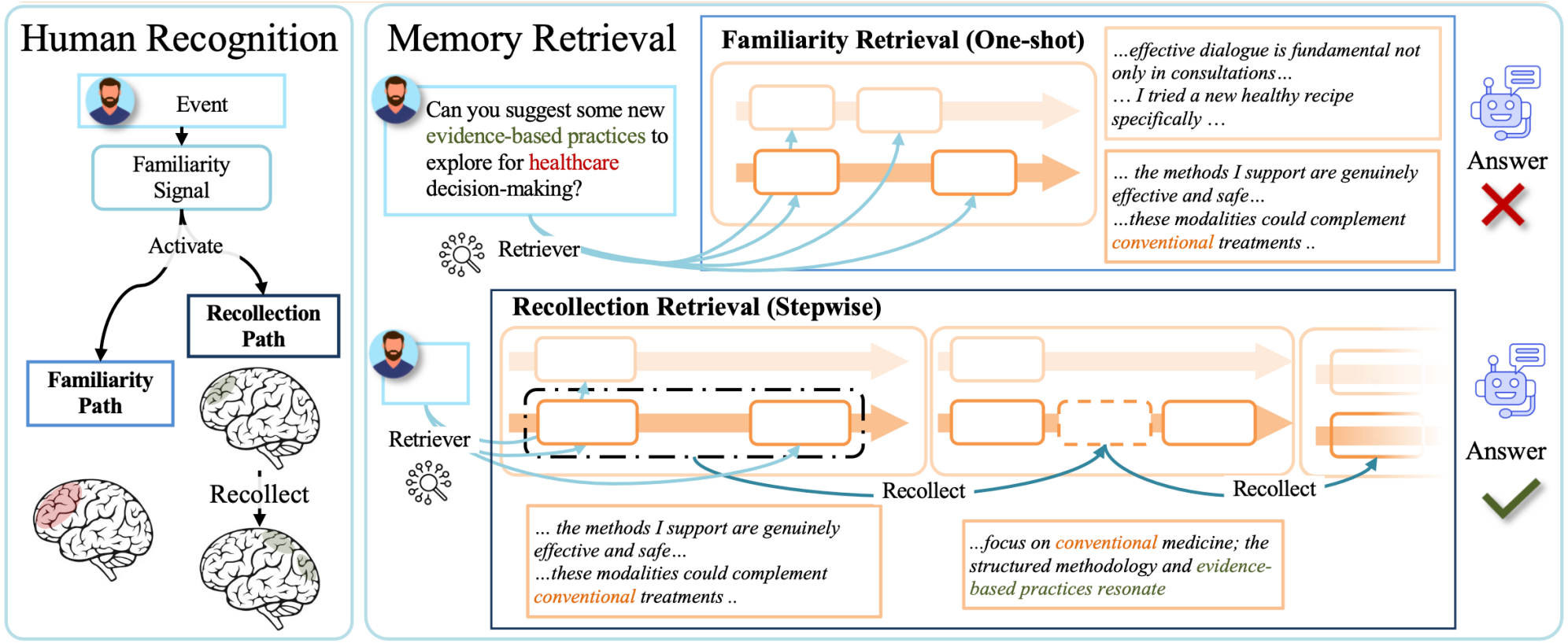
Case Study





1. **Compared with one-shot familiarity retrieval that returns fragmented memories, the recollection path progressively reconstructs temporally dispersed, semantically complementary evidence into a coherent memory trace for more grounded personalization.**

Content



Conclusion

1. We ground the design of personalized memory retrieval in the Recollection–Familiarity dual-process theory, formulating retrieval as a coordination of Familiarity and Recollection paths.
2. We introduce familiarity uncertainty-driven selection for adaptive switching between Familiarity and Recollection.
3. We develop a recollection retrieval based on clustering and query–centroid mixing, achieving chain-like evidence reconstruction only in embedding space.
4. RF-Mem is lightweight, relying solely on vector search and small-scale clustering, achieving high accuracy and recall at near one-shot retrieval latency.

Code

AML Group

Yingyi's Homepage

yingyizhang@mail.dlut.edu.cn

Thanks