

Long-Term Memory & LLMs

2024 Winter Seminar

인간중심컴퓨팅연구실 박사과정 류정우

Long-term Memory for LLMs

Intro



r/LocalLLaMA • 6 mo. ago
Davde_pq



Ai personal assistant with long term memory?

[Question](#) | [Help](#)



I want to train a personal model that would have access to my personal information, such as my calendar, and that could help me organise my life a bit.

I was initially thinking of using a small model (7/13b), and connect it to a vectordb for the long-term memory part, but I'm pretty new in the llm world and would like your guys advice.

Ideally the model should be able to:

- Remember my calendar and be able to access and modify the schedules.
- Answer questions about stored data. (Information retrieval).
- Deduce the importance of an event. (Going out for a drink < uni deadline).

I'd love to hear your guys opinions on this, and thank you in advance for any suggestions :)

 25 

 18

LLMs without Long-term Memory

Despite the remarkable abilities of LLMs,

- Inability to process **exceptionally lengthy inputs**
 - e.g. medical AI assistants - not to forget important **disease symptoms**
- Lack of long-term memory, an essential aspect of **human-like communication**
 - e.g. personal AI companions - need to recall past conversations **for rapport building**
- 보다 정확하고 안정적인 상호작용을 위해서는 long-term capabilities 갖춘 AI 시스템을 개발하는 것이 필요!

CLOVA CareCall

AI기반 전화 돌봄 서비스

출처: [Naver Cloud Platform](#)

국내 최초 초대규모 AI 안부 확인 서비스

한국어에 특화된 AI가 돌봄 대상자와 자연스러운 대화를 나누며 안부를 확인하고 상태를 기록합니다.

공감과 위로로 감동을 주는 AI 안부 전화

CLOVA CareCall은 돌봄이 필요한 대상자에게 AI가 주기적으로 전화를 걸어 건강, 식사, 수면 등과 관련된 일상적인 안부를 묻고 이상 징후를 모니터링하는 AI 안부 확인 서비스입니다.

CLOVA CareCall의 AI는 세계 최대 한국어 언어 모델인 HyperCLOVA를 적용하여 실제 사람과 대화하듯 다양한 주제로 자연스럽게 상호작용할 수 있습니다. 또한 부드럽고 친절한 목소리로 돌봄 대상자에게 관심, 공감, 위로 등의 감정을 담아 표현하는 말벗이 될 수 있습니다.



[대화의 기술, 사람을 돕다] 클로바 케어콜 인터뷰 - 연극배우 손숙

기억하기

이전에 주고받았던 대화 내용을 다음 통화에 활용하는 대화 기술로 AI 대화형 서비스에는 최초로 도입한 기능입니다. 모든 대상자와 똑같은 대화를 나누는 게 아니라 대상자에 따라 개인화된 대화를 나눔으로써 대화 만족도를 높이고 친밀감을 형성하여 대상자의 상태를 지속해서 관찰할 수 있습니다.



Paper #1

MemoryBank

Paper 1

MemoryBank

MemoryBank: Enhancing Large Language Models with Long-Term Memory

Wanjun Zhong¹, Lianghong Guo¹, Qiqi Gao², He Ye³, Yanlin Wang¹

¹ Sun Yat-Sen University ² Harbin Institute of Technology

³ KTH Royal Institute of Technology

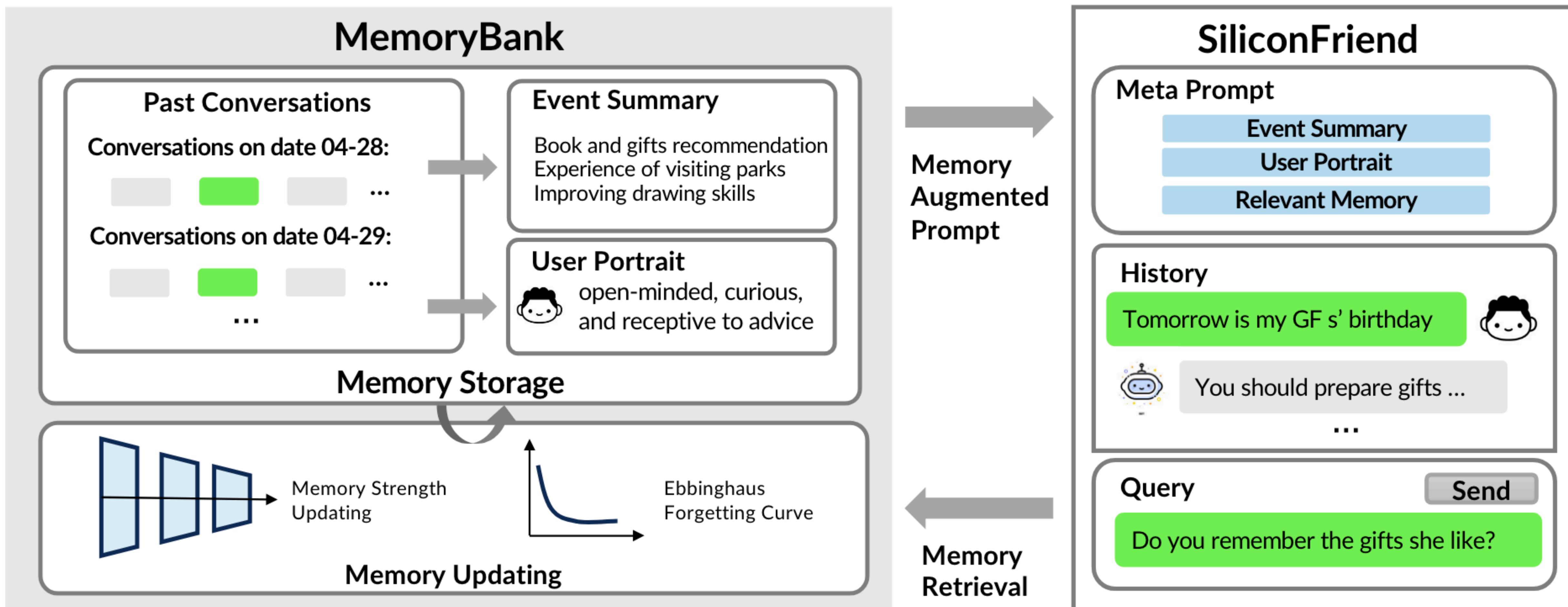
{zhongwj25@mail2, wangylin36@mail}.sysu.edu.com

2231612405@qq.com, 18b903026@stu.hit.edu.cn

heye@kth.se

Overview

MemoryBank



Memory Mechanism Tailored for LLMs

MemoryBank

- Memory Storage
 - **In-depth Memory Storage:** each piece of dialogue + timestamp
 - **Hierarchical Event Summary:** processes and distills conversations into a high-level summary of **daily events**
 - **Dynamic Personality Understanding:** continuously assesses and updates these understandings with the long-term interactions and creates daily personality insights

Memory Mechanism Tailored for LLMs

MemoryBank

- Memory Updating Mechanism: 덜 중요하거나 덜 불러온 기억은 잊게 하는 것 (inspired by *Ebbinghaus Forgetting Curve* theory)
 - **Rate of Forgetting:** 시간이 흐름에 따라 memory retention이 감소
 - **Time and Memory Decay:** 처음에는 급격하게, 이후에는 완만하게 잊어버림
 - **Spacing Effect:** 다시 습득하는 지식은 처음에 비해 받아들이기 쉬움
- 망각 곡선 이론을 실제 모델에 반영한 것이 재미있는 포인트

SiliconFriend: the System Development

MemoryBank

- 2-Stage Development
 1. **LoRA tuning with psychological dialogue data:** human companion처럼 공감해주고 정서적 지원을 하도록 만들기 위한 단계
 2. **MemoryBank Integration:** 과거의 상호작용을 바탕으로 기억을 저장하고, 업데이트하는 능력을 탑재하기 위한 단계

SiliconFriend: Psychological Companion

MemoryBank

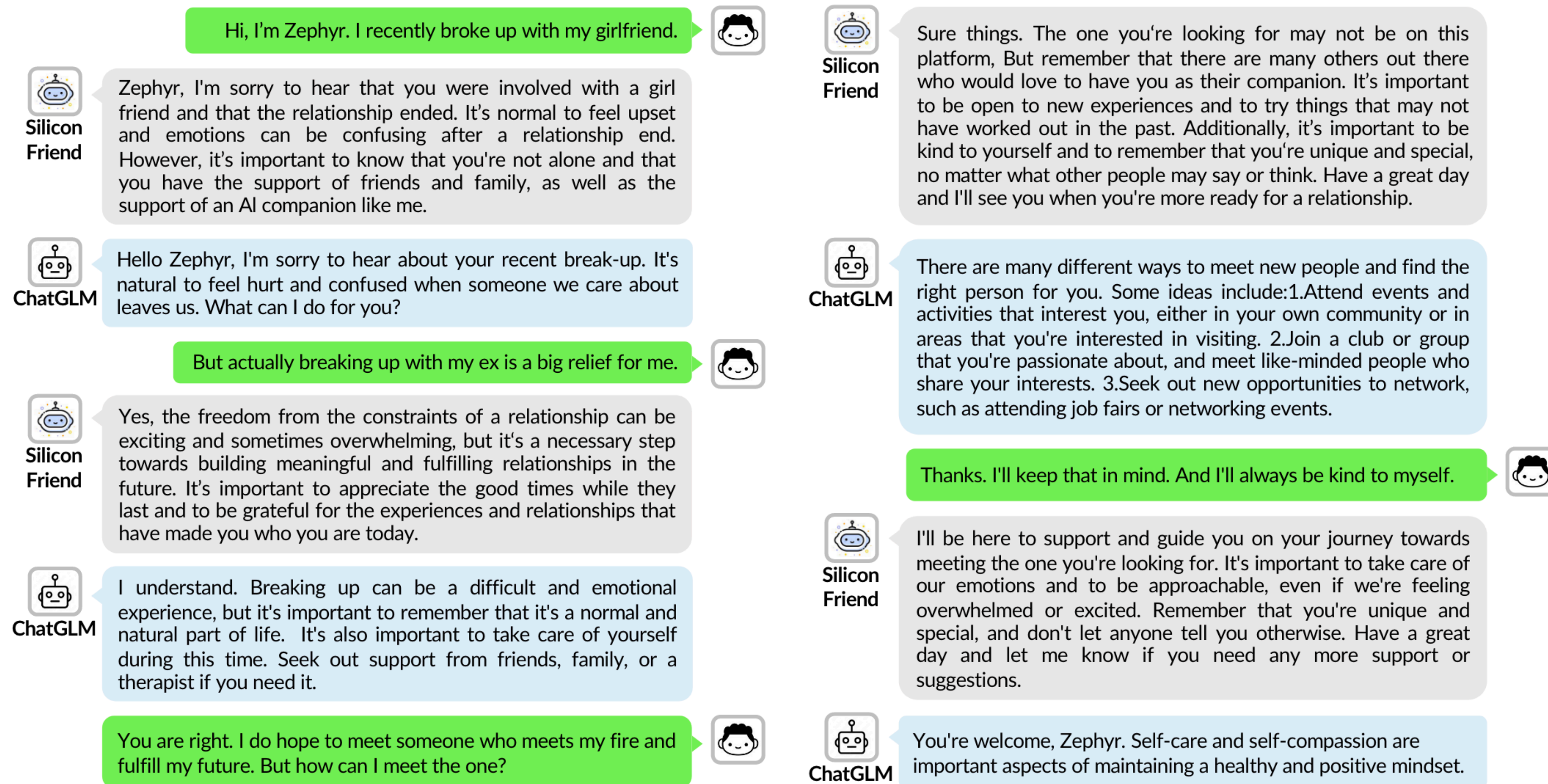


Figure 2: Example of consulting SiliconFriend_{ChatGLM} for psychological companionship. Overall, SiliconFriend can provide more empathetic response, offer constructive emotional support to user and help him to face sorrow with positive attitude.

SiliconFriend: Memory Recall

MemoryBank

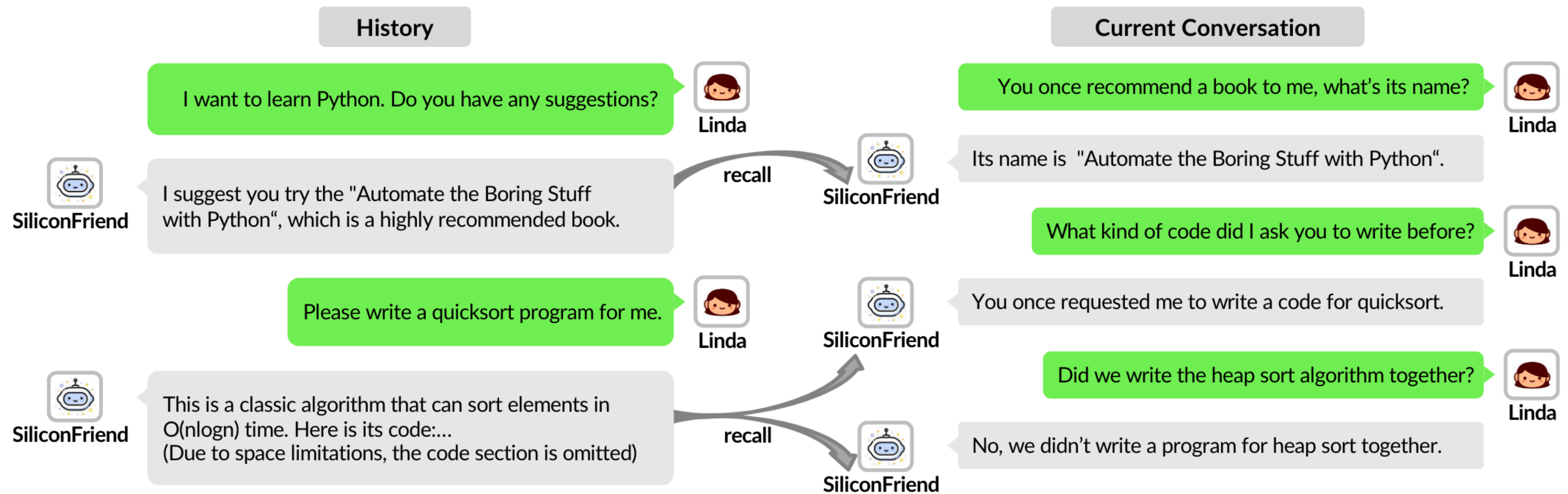


Figure 3: Example responses from SiliconFriend _{BELLE} in memory recall.

SiliconFriend: Understanding Personalities

MemoryBank

Linda is an introverted, determined, ambitious, and responsible girl who values personal growth, exploring new cultures and hobbies, and seeking advice from others.

What do you know about me?



Linda



Silicon Friend

Based on our interactions and your past conversations, I know that you are an introverted person who values personal growth, determination, ambition, and responsibility. You have a keen interest in exploring new cultures, trying out new hobbies, and seeking advice from others. Is there anything else about you that you would like me to know?

I'm considering engaging in some activities over the weekend. Could you provide some suggestions?



Linda



Silicon Friend

Absolutely! Based on your interests in exploring new cultures and trying out new hobbies, here are some suggestions for you:

1. Attend a cooking class to learn how to make dishes from a different cuisine.
2. Visit a museum or art exhibition to explore a new culture or gain insights into a different art form.

Emily is an open-minded and curious girl with lots of interests. While she is receptive to advice, she sometimes experiences self-doubt and uncertainty about her future.

What do you know about me?



Emily



Silicon Friend

You seem open-minded and receptive to advice. If you'd like to share more about your interests or anything else, feel free to let me know and I'll do my best to help!

I'm considering engaging in some activities over the weekend. Could you provide some suggestions?



Emily



Silicon Friend

Of course! It's great that you're interested in doing something this weekend. Depending on what you enjoy, there are many fun and engaging activities to do. For example, if you're interested in the outdoors, you might consider going hiking or taking a bike ride. If music is more your thing, check out local concerts or learn a new instrument. Ultimately, it's important to find something that you enjoy and that also allows you to recharge and relax. Whatever you decide to do, make sure it aligns with your interests and leaves you feeling refreshed!

Figure 4: Example responses from SiliconFriend_{ChatGPT} to users with different personalities.

Quantitative Analysis

MemoryBank

- **Retrieval Accuracy** evaluates whether the relevant memory is successfully recalled (labels: {0: no; 1: yes}).
- **Response Correctness** evaluates if correctly answering the probing question (labels: {0: wrong; 0.5: partial; 1 : correct}).
- **Contextual Coherence** evaluates whether the response is naturally and coherently generated, *e.g.*, connecting the dialogue context and retrieved memory (labels: {0: not coherent; 0.5: partially coherent; 1: coherent}).

Table 2: Results of quantitative analysis.

Language	Model	Retrieval Acc.	Correctness	Coherence	Ranking
English	SiliconFriend _{ChatGLM}	0.809	0.438	0.68	0.498
	SiliconFriend _{BELLE}	0.814	0.479	0.582	0.517
	SiliconFriend _{ChatGPT}	0.763	0.716	0.912	0.818
Chinese	SiliconFriend _{ChatGLM}	0.84	0.418	0.428	0.51
	SiliconFriend _{BELLE}	0.856	0.603	0.562	0.565
	SiliconFriend _{ChatGPT}	0.711	0.655	0.675	0.758

Contributions

MemoryBank

- MemoryBank: 인간과 유사한 형태의 long-term memory mechanism(store, recall, update memory, and draw user portrait)
- SiliconFriend: MemoryBank를 탑재하고 심리적 대화로 조정된 LLM 기반 AI 동반자 SiliconFriend를 통해 MemoryBank의 실제 적용 가능성을 보여줍니다
- 일반화 가능성 제시: 1) 오픈 소스 및 비공개 LLM 모두 수용, 2) bilingual(Chinese & English), 3) applicability with and without memory forgetting mechanism

Paper #2

Think-in-Memory

Paper 2

Think-in-Memory

Think-in-Memory: Recalling and Post-thinking Enable LLMs with Long-Term Memory

Lei Liu*

liulei1497@gmail.com

CUHK-Shenzhen, Ant Group

Xiaoyan Yang

joyce.yxy@antgroup.com

Ant Group

Yue Shen[†]

zhanying@antgroup.com

Ant Group

Binbin Hu, Zhiqiang Zhang

{bin.hbb,lingyao.zzq}@antfin.com

Ant Group

Jinjie Gu

jinjie.gjj@antfin.com

Ant Group

Guannan Zhang

zgn138592@antfin.com

Ant Group

Framework Overview

Think-in-Memory

- Main components
 - Each conversation turn: (user's query **Q**, agent's response **R**)
 - Agent, **A**: a pre-trained LLM to facilitate dynamic conversations
 - Memory Cache, **M**: a continually growing hash table of key-value pairs, where key is the hash index and **value is a single thought**
 - Hash-based Mapping, **F**: to quickly save and find the relevant thoughts in **M**

Workflow

Think-in-Memory

- Stage-1: Recall & Generation
 - 사용자로부터 주어진 **Q**에 대해 agent **A**가 relevant thoughts(self-generated reasoning)를 추출하여 답하는 과정 - 직접 **recall**할 수 있는 형태로 **external memory**에 저장
- **Stage-2: Post-think and Update(차별점)**
 - Stage-1에서 답변하고난 이후, agent **A**에게 **Q-R** 쌍에 대해 **사후사고(post-think)**를 하도록 하여 **기존의 기억 M**을 최신화

Memory Updating - Insert, Forget, Merge

Think-in-Memory

Prompt for Generating Thoughts

Given the following question and response pairs, please extract the relation (subject, relation, object) with corresponding text:

Example 1.

Input:

Question: Do you have any company recommendations for me?

Response: I recommend Google.

Output:

(Company, Recommended, Google).

Recommended company is Google.

Example 2.

Input:

Question: Which City is the capital of China?

Response: Beijing.

Output:

(China, Capital, Beijing).

The capital of China is Beijing.

Input:

Question: Do you have any book recommendations for me?

Response: I recommend "The Little Prince".

Output:

Prompt for Forgetting Thoughts

Given the following thoughts, please remove the counterfactual thoughts or contradictory thoughts:

Example 1.

Input:

The capital of China is Beijing.

The capital of China is Shanghai.

The capital of the United States is Washington.

The capital of the United States is New York.

Output:

The capital of China is Beijing.

The capital of the United States is Washington.

Example 2.

Input:

Michael likes to play football.

Michael does not like to play football.

James likes to swim.

Mary likes to read books.

Output:

James likes to swim.

Mary likes to read books.

Input:

[A group of thoughts]

Output:

Prompt for Merging Thoughts

Given the following thoughts, please merge the similar thoughts with the same entity:

Example 1.

Input:

John works as an actor.

John works as a director.

John works as a writer.

Mike works as a teacher.

Output:

John works as an actor, a director, and a writer.

Mike works as a teacher.

Example 2.

Input:

Michael likes to play football.

Michael likes to play basketball.

James likes to swim.

Mary likes to read books.

Output:

Michael likes to play football and basketball.

James likes to swim.

Mary likes to read books.

Input:

[A group of thoughts]

Output:

Figure 3: An example of prompts for generating thoughts.

Figure 4: An example of prompts for forgetting thoughts.

Figure 5: An example of prompts for merging thoughts.

Memory Updating

Think-in-Memory

Prompt for Generating Thoughts

Given the following question and response pairs, please extract the relation (subject, relation, object) with corresponding text:

Example 1.

Input:

Question: Do you have any company recommendations for me?
Response: I recommend Google.

Output:

(Company, Recommended, Google).
Recommended company is Google.

Example 2.

Input:

Question: Which City is the capital of China?
Response: Beijing.

Output:

(China, Capital, Beijing).
The capital of China is Beijing.

Input:

Question: Do you have any book recommendations for me?
Response: I recommend "The Little Prince".

Output:

Prompt for Forgetting Thoughts

Given the following thoughts, please remove the counterfactual thoughts or contradictory thoughts:

Example 1.

Input:

The capital of China is Beijing.
The capital of China is Shanghai.
The capital of the United States is Washington.
The capital of the United States is New York.

Output:

The capital of China is Beijing.
The capital of the United States is Washington.

Example 2.

Input:

Michael likes to play football.
Michael does not like to play football.
James likes to swim.
Mary likes to read books.

Output:

James likes to swim.
Mary likes to read books.

Input:

[A group of thoughts]

Output:

Prompt for Merging Thoughts

Given the following thoughts, please merge the similar thoughts with the same entity:

Example 1.

Input:

John works as an actor.
John works as a director.
John works as a writer.
Mike works as a teacher.

Output:

John works as an actor, a director, and a writer.
Mike works as a teacher.

Example 2.

Input:

Michael likes to play football.
Michael likes to play basketball.
James likes to swim.
Mary likes to read books.

Output:

Michael likes to play football and basketball.
James likes to swim.
Mary likes to read books.

Input:

[A group of thoughts]

Output:

Fine-tuning with LORA

Figure 3: An example of prompts for generating thoughts.

Figure 4: An example of prompts for forgetting thoughts.

Figure 5: An example of prompts for merging thoughts.

Results

Think-in-Memory

Table 2: Comparison Results on Three Datasets. Top-5 thoughts are recalled from the memory cache.

Dataset	LLM	Language/Topic	Memory	Retrieval Accuracy	Response Correctness	Contextual Coherence
GVD	ChatGLM	English/Open	SiliconFriend	0.809	0.438	0.680
			TiM (Ours)	0.820	0.450	0.735
		Chinese/Open	SiliconFriend	0.840	0.418	0.428
			TiM (Ours)	0.850	0.605	0.665
Kdconv	ChatGLM	Chinese/Film	✗	-	0.657	0.923
			TiM (Ours)	0.920	0.827	0.943
		Chinese/Music	✗	-	0.666	0.910
			TiM (Ours)	0.970	0.826	0.926
		Chinese/Travel	✗	-	0.735	0.906
			TiM (Ours)	0.940	0.766	0.912
	Baichuan2	Chinese/Film	✗	-	0.360	0.413
			TiM (Ours)	0.913	0.743	0.870
		Chinese/Music	✗	-	0.253	0.283
			TiM (Ours)	0.900	0.710	0.780
		Chinese/Travel	✗	-	0.207	0.280
			TiM (Ours)	0.833	0.757	0.807
RMD	ChatGLM	Chinese/Medical	✗	-	0.806	0.893
			TiM (Ours)	0.900	0.843	0.943
	Baichuan2	Chinese/Medical	✗	-	0.506	0.538
			TiM (Ours)	0.873	0.538	0.663

- **Retrieval Accuracy** evaluates whether the relevant memory is successfully recalled (labels: {0: no; 1: yes}).
- **Response Correctness** evaluates if correctly answering the probing question (labels: {0: wrong; 0.5: partial; 1 : correct}).
- **Contextual Coherence** evaluates whether the response is naturally and coherently generated, *e.g.*, connecting the dialogue context and retrieved memory (labels: {0: not coherent; 0.5: partially coherent; 1: coherent}).

Long-term Memory Mechanism

Think-in-Memory

Conversation History

User: Do you have any book recommendations for me?

Agent: I recommend "The Little Prince".

User: Thank you. How about this book?

Agent: This book is very interesting and full of emotions.

⋮

User: Do you have any movie recommendations for me?

Agent: You can go see "The Wandering Earth".

⋮

User: Do you have any song recommendations for me?

Agent: I recommend "Listen to Mom".

User: You recommended a movie before, how about it?

Agent: It is "The Wandering Earth". This movie is with stunning visuals.

User: 영화 하나 추천해줄래?

Agent: "The Wandering Earth" 한 번 봐봐.

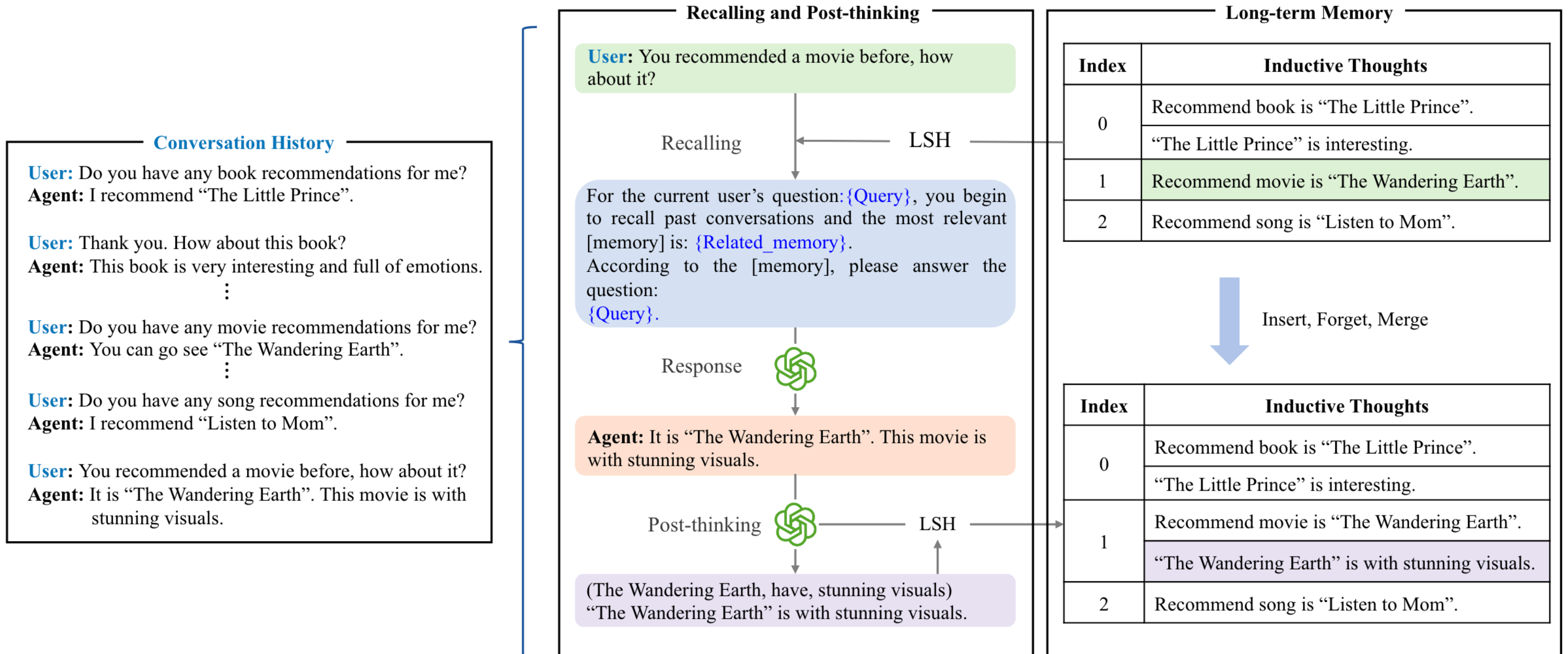
시간이 흐르고...

User: 지난 번에 추천해준 영화 있잖아, 그거 어때?

Agent: "The Wandering Earth". 이 영화 비주얼이 끝내줘.

Long-term Memory Mechanism

Think-in-Memory



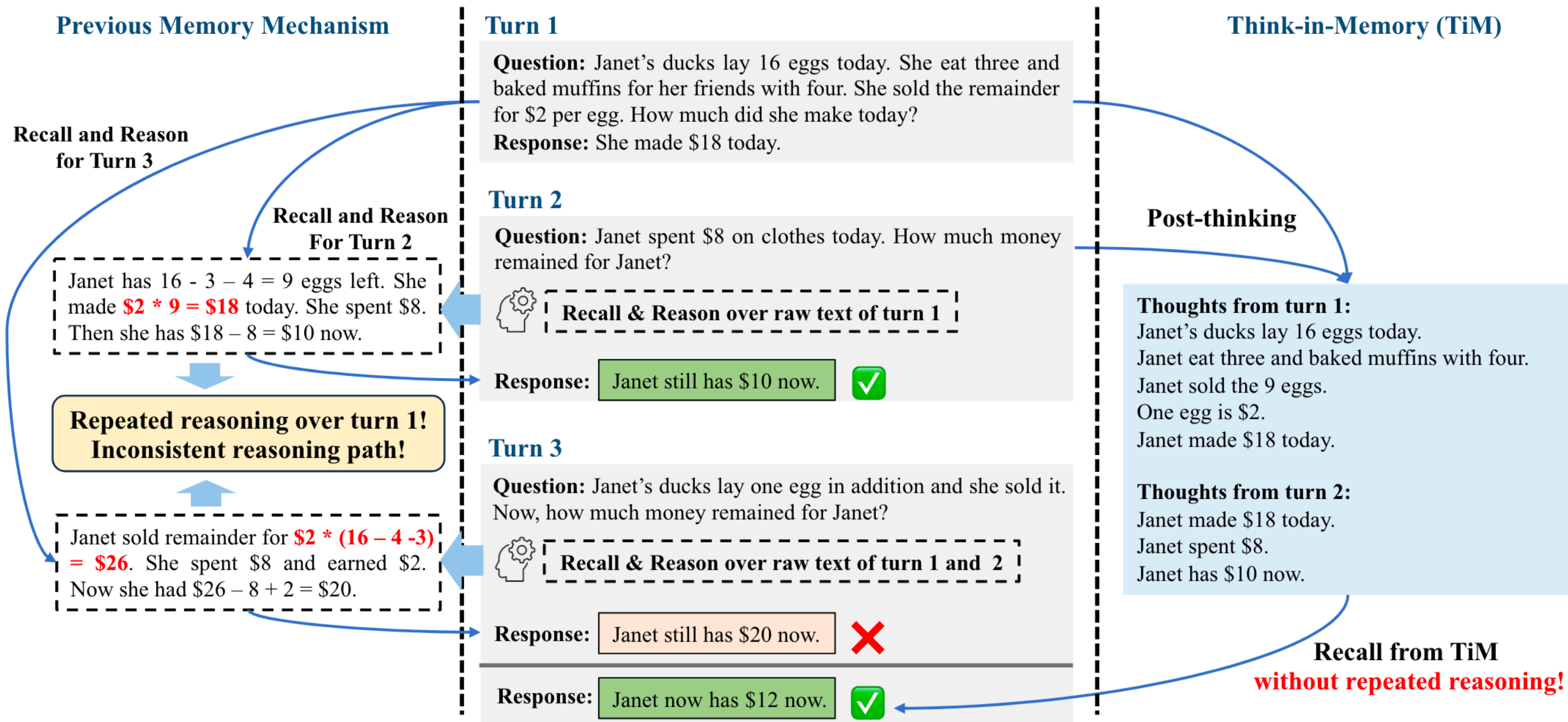


Figure 1: Comparisons between previous memory mechanisms with our proposed TiM. (Left): Existing memory mechanisms mainly save raw text of previous turns, which require repeated reasoning over the same history. This easily leads to the inconsistent reasoning path (*i.e.*, red part of the left) with wrong response. (Right): The proposed TiM stores the thoughts of LLMs for previous turns, which can avoid such inconsistency without repeated reasoning (*i.e.*, red part of the right).

Contributions

Think-in-Memory

- TiM(Think-in-Memory): 인간과 유사한 형태의 long-term memory mechanism 구현
- Human-like 생각 정리: 인간의 인지 과정을 반영하여 기억 속의 생각을 정리하기 위한 몇 가지 기본 원칙을 공식화
- 광범위한 실험 진행: 1) open ~ specific domains, 2) bilingual(Chinese & English), 3) correctness & coherence improvement

contact me

jeongwoo@snu.ac.kr

Background for TiM-LLM Application

Insert, Forget, Merge Operations

Example:
Merge two thoughts

User consults symptoms.
Agent analyzes symptoms.
Agent provides treatments and suggestions.

Hash Index	Thoughts
1	User consults symptoms.
1	Agent analyzes symptoms.
1	Agent provides suggestions.
1	Agent provides treatments.
2	Symptom includes red spots on the tongue.
2	Symptom includes a tingling sensation.
2	Symptom includes grain-like white spots on gums and throat.
3	Diagnosis is inflammation of the oral mucosa.
4	Suggestion is to go to a specialized oral clinic for examination.
4	Suggestion is to inform the doctor for an accurate diagnosis.
⋮	⋮

Incomplete response without TiM.



Oral mucosal inflammation may be one of the causes of these symptoms. However, I cannot be certain if it is the sole cause, as further examinations are needed to make an accurate diagnosis.
Regarding your use of cefuroxime, it generally does not have a direct impact on oral mucosal inflammation.

User Interface

TiM-LLM Application

Hello! I'm here to consult about my symptoms.

Hello! Please describe your symptoms in detail. I will analyze and diagnose your condition, and provide reasonable treatment plans or suggestions for you!

Background Process: Post-thinking!

I am feeling uncomfortable. I have red spots on my tongue and a tingling sensation. There are many grain-like white spots on my gums and throat. What is this disease?

Background Process: Recalling!

Based on your symptom description, there may be inflammation of the oral mucosa. I suggest you go to a specialized oral clinic for an examination and inform the doctor about your symptoms and severity so that the doctor can make a correct diagnosis.

Background Process: Post-thinking!

⋮ Multiple rounds

Background Process: Recalling!

Is oral mucosal inflammation the only reason for these symptoms? I have used cefuroxime before, would it affect the current symptoms?

Oral mucosal inflammation is not the only cause of these symptoms. Other factors such as fungal infections, bacterial infections, allergies may also cause these symptoms. If cefuroxime was used for a bacterial infection, it may provide some assistance in treatment, but it can also disrupt the beneficial bacterial flora in the oral cavity. It is best to seek examination from a professional doctor to obtain an accurate diagnosis.