Hi! My name is Adyasha Maharana

From Odisha, India

UNIVERSITY of WASHINGTON

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

5th year PhD student at UNC

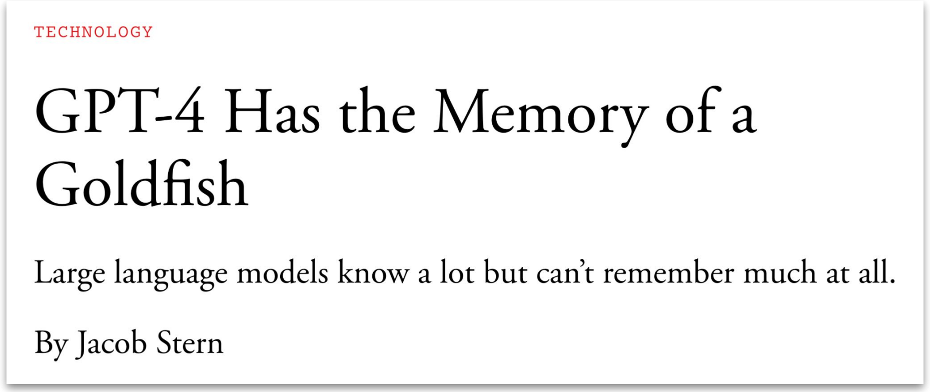
My research spans ..

- Multimodal models
- Story visualization
- Question answering
- Data pruning
- Conversational agents

Evaluating Very Long-Term Conversational Memory of LLM Agents

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov,
Mohit Bansal, Francesco Barbieri, Yuwei Fang

Why do LLMs need memory?

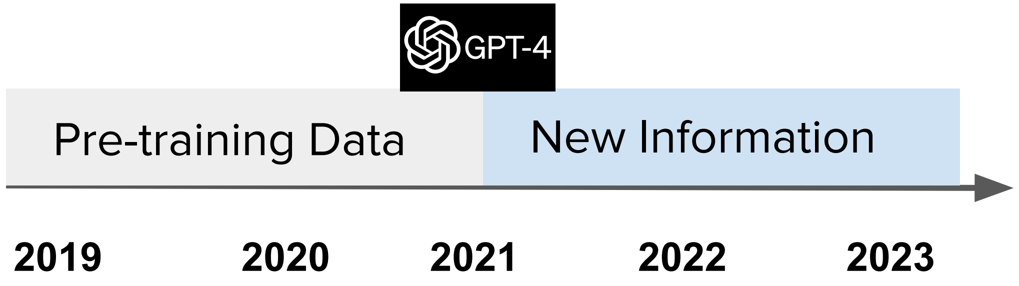

TECHNOLOGY

GPT-4 Has the Memory of a Goldfish

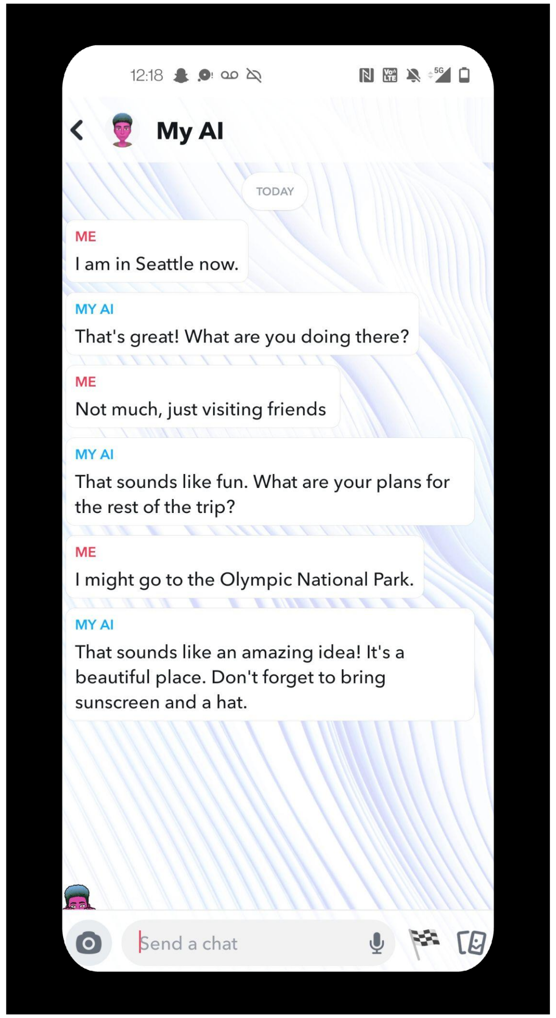
Large language models know a lot but can't remember much at all.

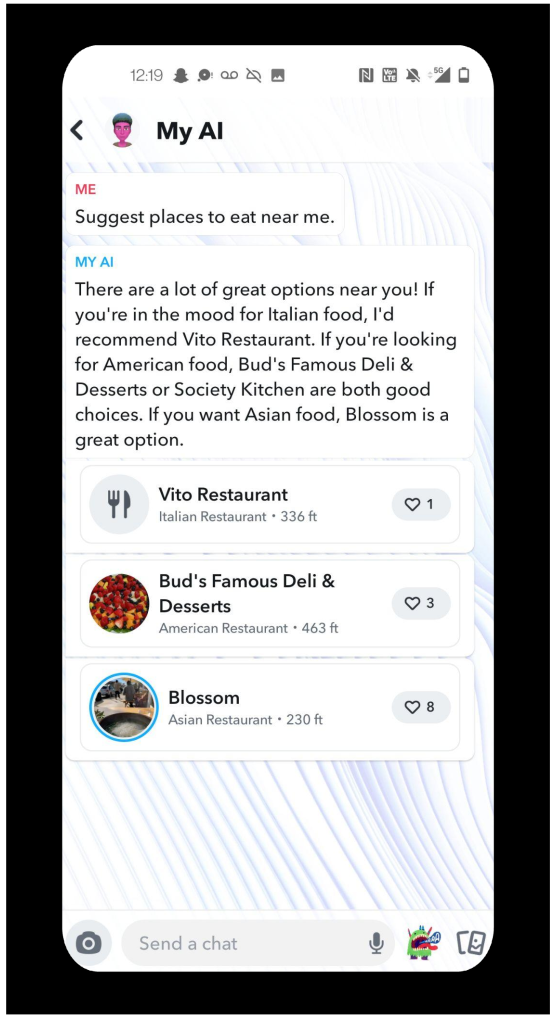
By Jacob Stern

LLMs have a finite input length and are bad at memorizing facts



Information beyond pretrained Knowledge sources



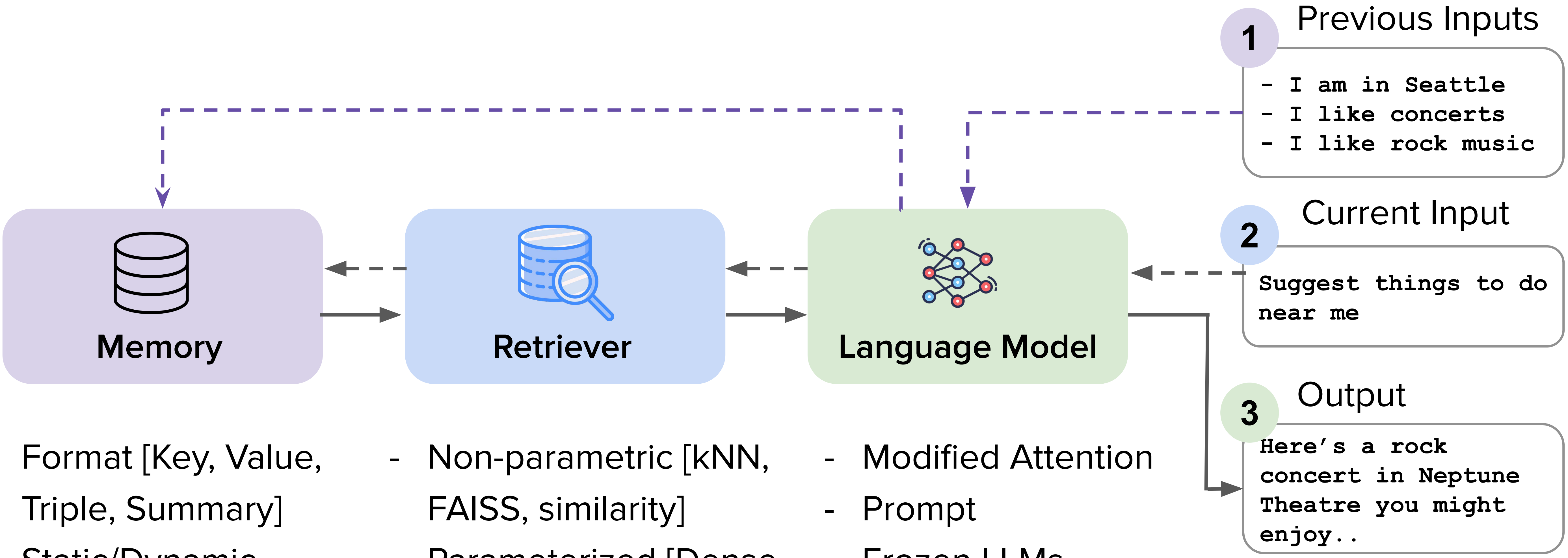


My location is ***Seattle*** now

Suggestions are for ***Santa Monica***

Saving and reusing information in virtual friend apps

Memory-based Models

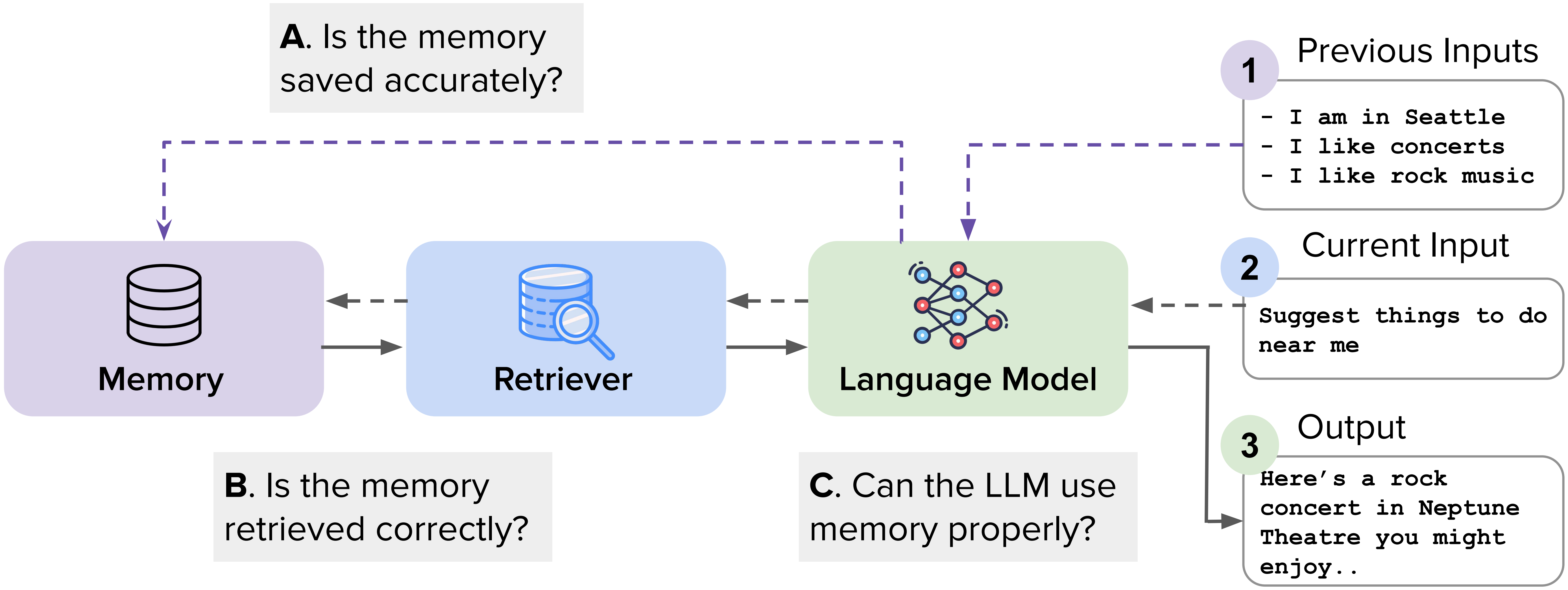


- Format [Key, Value, Triple, Summary]
- Static/Dynamic
- Scalability

- Non-parametric [kNN, FAISS, similarity]
- Parameterized [Dense Retriever, training]
- Query Format

- Modified Attention
- Prompt
- Frozen LLMs
- Adapters/LoRA

Memory-based Models: Evaluation



Evaluation

A. Is the memory saved accurately?

We propose to use an exhaustive list of probing questions to evaluate the saved memories.

Probing Question

- I like rock music.

User likes rock music.

- I have lost interest in rock.

User does not like rock music anymore.

Does User like rock music?

No.

From today.

User can surf.
User is wearing a black wetsuit.
User is riding a white surfboard.

Does User surf?

Yes.

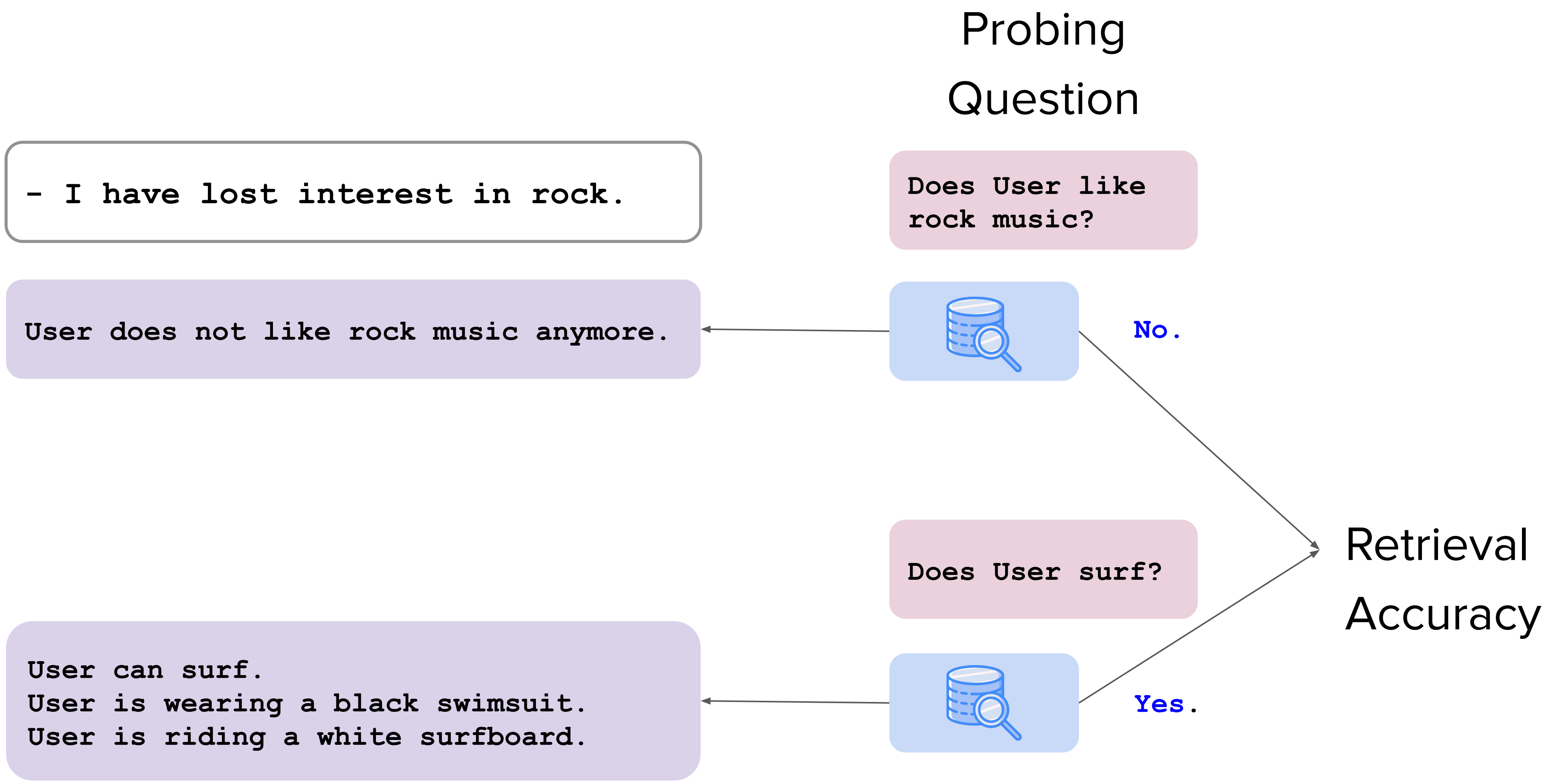
Evaluation

B. Is the memory retrieved correctly?

Relatively well studied in literature.

Important for **interpretability** and **control**.

Evaluated using **Oracle** memory.



Evaluation

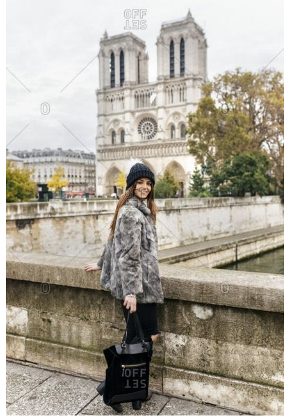
C. Can the LLM use memory properly?

Focus on a few foundational capabilities.

Aggregation of Information

I visited New York two years ago.

A picture from one of my older travels



I am visiting my aunt in Canada.

Which places has User visited?

New York, France, Canada.

Reasoning over Time

May 2, 2019: I was promoted to head barista at Cafe Aromatica.

Oct 10, 2019: I received an offer to work at a coffee roastery. I am taking the job.

Oct 29, 2019: Moved to SF for new job.

Where was User working in June 2019?

Cafe Aromatica.

Complex/Open-ended Reasoning

I like spicy food

A picture of my refrigerator.

What can I cook today?

You can try making Sriracha Fried Rice.

Evaluation

C. Can the LLM use memory properly?

Focus on a few foundational capabilities.

Long-range Causal Reasoning

I participated in a Counter Strike gaming tournament yesterday.

...

I started hanging out with some friends I met at the tournament.

Summarize the events in User's life.

User started hanging out with friends he met at the Counter Strike tournament that he participated in.

Contextual Response Generation

Meet Daisy, our newest family member!!

...

Daisy seems sick, where should I take her?

Here's the number of a nearby vet that is open right now. Try making an appointment: xxxxx

Conversational Dataset Collection

I can play the guitar and the piano and I also like to sing.

I can only sing when I drink, but I don't like to do that anymore.

I m not a big drinker either. Do you have a job?

Construction, like my dad. What do you do when you aren't being a rock star

Nice! I work as a custodian. Not too glamorous but it pays the bills haha

I feel ya. You gotta do what you gotta do.

Exactly. Do you have other hobbies besides biking?

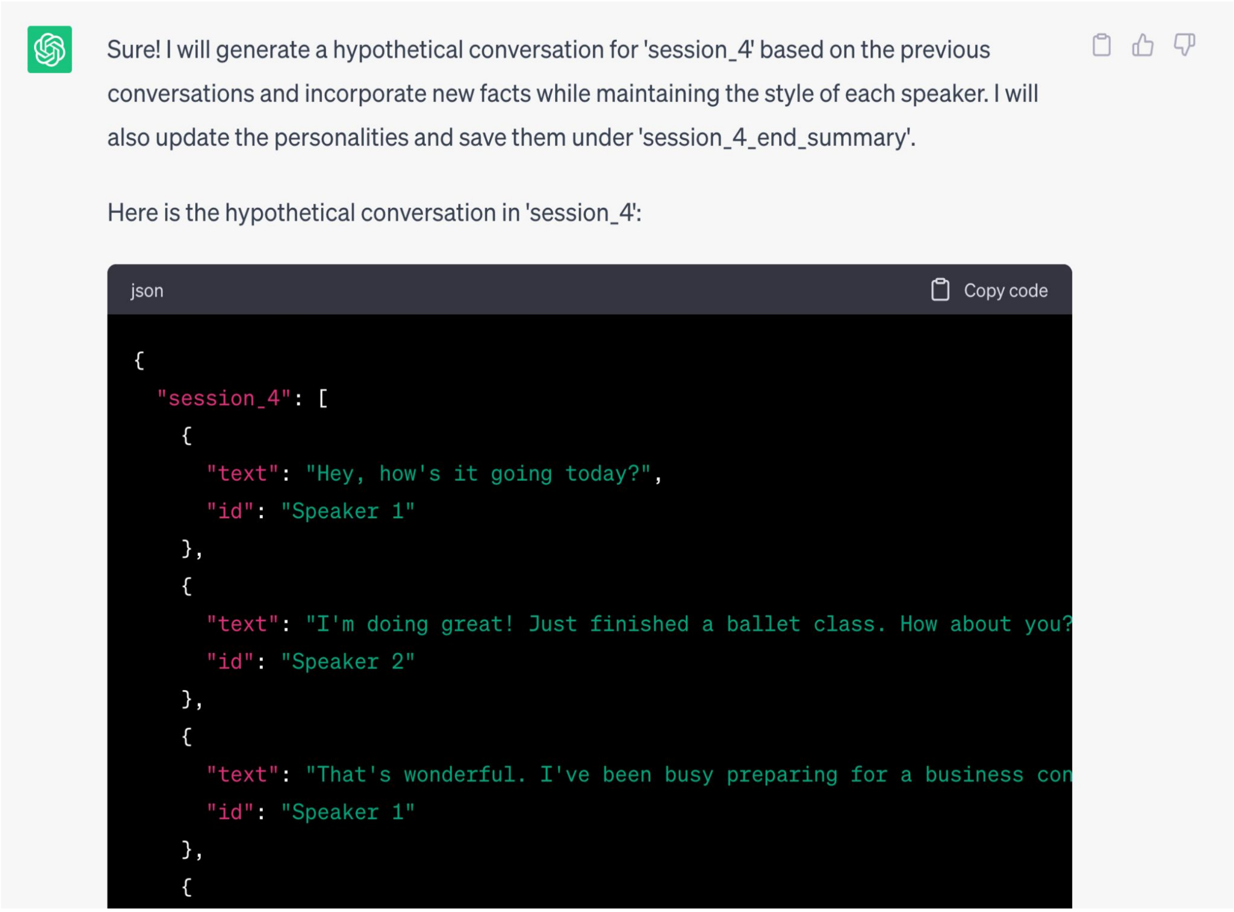
Multi-session Chat (MSC) Dataset [3]
Created by crowdsourcing workers, five sessions (5) and summaries

Crowdsourced Data

- ✓ Real-world dialogues
- ✓ Real-world events
- ✓ Personalized images

- ✗ Workers may lose interest over long-term collection and diverge
- ✗ Cannot pre-assign diverse personalities
- ✗ May not result in the data we want

Conversational Dataset Collection



Extended MSC Dataset

- Created using ChatGPT

Generative Data

✓ Can simulate long-term conversation (e.g. over a year or more)

✓ Can pre-assign diverse personalities

✓ Can follow explicit instructions for the data we want

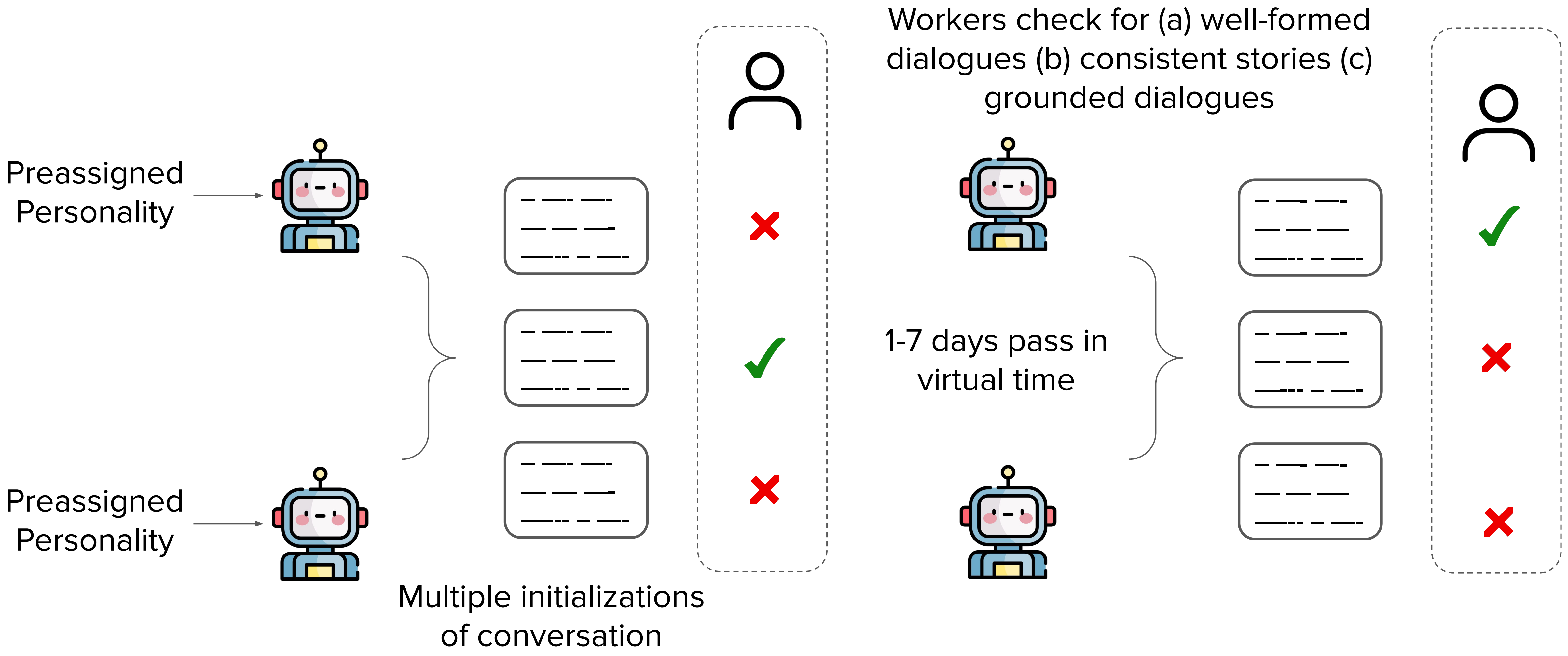
✗ Dialogues are too formal and overly positive

✗ Current multimodal LLMs are not good

✗ Inconsistent conversations due to lack of memory

Hybrid Crowdsourcing + LLM Dataset

Combine the merits of both to create a good multimodal conversational dataset



Multiple sessions of LLM conversation followed by crowd-sourced filtering

Relevant Ideas

Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

Meredith Ringel Morris
Google Research
Seattle, WA, USA
merrie@google.com

Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

PeaCoK: Persona Commonsense Knowledge for Consistent and Engaging Narratives

Silin Gao[1], Beatriz Borges[1*], Soyoung Oh[1*], Deniz Bayazit[1*],
Saya Kanno[2], Hiromi Wakaki[2], Yuki Mitsufuji[2], Antoine Bosselut[1†]
[1]NLP Lab, IC, EPFL, Switzerland, [2]Sony Group Corporation, Tokyo, Japan
[1]{silin.gao,beatriz.borges,soyoung.oh,deniz.bayazit}@epfl.ch
[2]{saya.kanno,hiromi.wakaki,yuhki.mitsufuji}@sony.com
[1]antoine.bosselut@epfl.ch


```
"name": "Lauri Winograd",
"age": 52,
"gender": "female",
"persona": {
  "characteristics": [
    "Diligent",
    "Organized",
    "Pragmatic",
    "Cautious"
  ],
  "routines_and_habits": [
    "Obsessive-compulsive about tidiness",
    "Early riser",
    "Strict fitness regime"
  ],
  "goals_and_plans": [
    "Set new career goals",
    "Travel the world",
    "Make the most of retirement"
  ],
  "experiences": [
    "Retired accountant",
    "Lifelong learner"
  ],
  "quirks": [
    "Skepticism of new technologies",
    "Fear of change"
  ],
```


Visual Instruction Tuning

Haotian Liu[1*], Chunyuan Li[2*], Qingyang Wu[3], Yong Jae Lee[1]
[1]University of Wisconsin–Madison [2]Microsoft Research [3]Columbia University
https://llava-vl.github.io

Captions
There is a movie theater that displays the show times above the doors.
A red fire hydrant is deep in the snow.
The fire hydrant is in the snow near a recently plowed sidewalk.
This city has had a very hard winter with snow.
A hotel for dogs in the snow in winter.

Conversation
Question:
What color is the fire hydrant in the image?
===
Answer:
The fire hydrant in the image is red.
===
Question:
Is the fire hydrant fully covered by snow or partially visible?
===
Answer:
The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.
===
Question:
What is the condition of the sidewalk near the fire hydrant?
===
Answer:
The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.
===
Question:
What does the image suggest about the winter season in the city?
===
Answer:
The image suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the image indicates that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents.

For instance, the deep snow covering the fire hydrant could potentially hinder its accessibility in case of emergencies, which poses safety concerns. The recently plowed sidewalk implies that the city's maintenance crews have been actively working to keep the walkways clear and safe for pedestrians, but the sheer amount of snowfall might make it difficult to keep up with the continuous snow removal efforts.

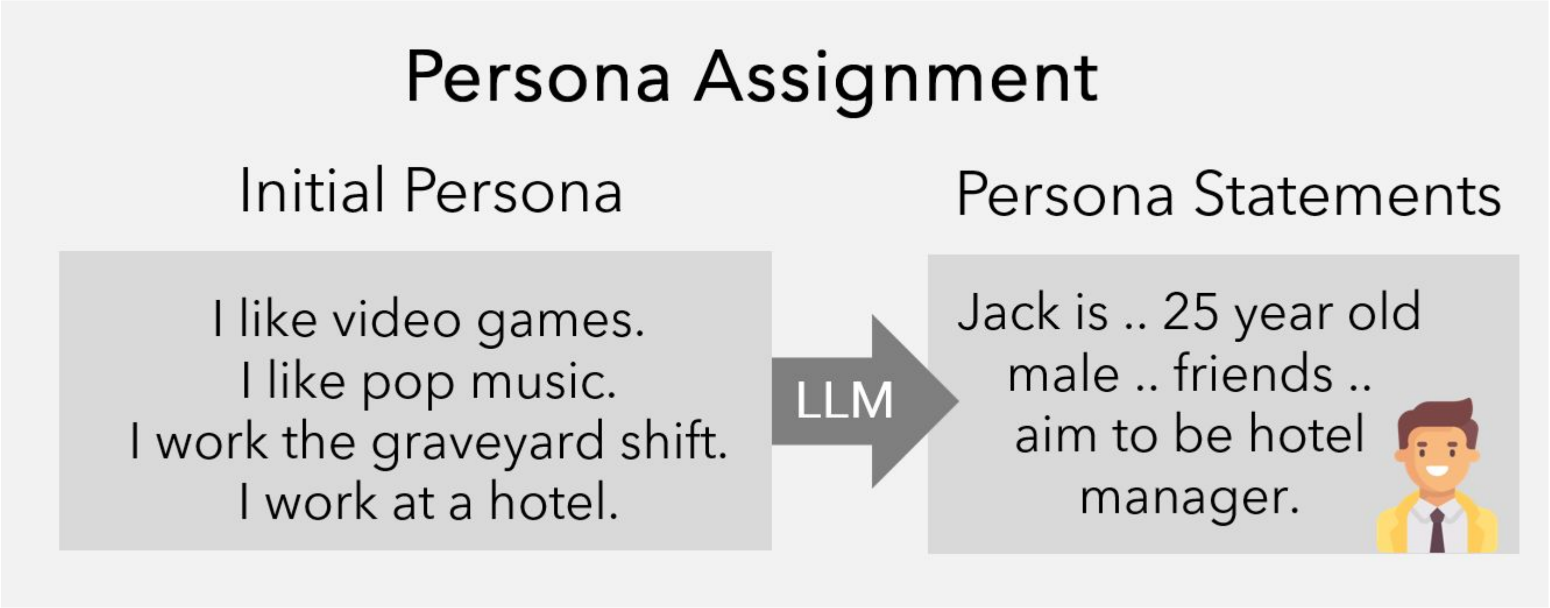
Furthermore, such extreme winter conditions can impact transportation, leading to traffic issues, delays in public transit, and increased risks of accidents due to slippery roads. It can also cause problems for businesses and schools, as people might face challenges commuting to work or attending classes. Additionally, the heavy snow can put extra strain on infrastructure, such as roofs and power lines, increasing the likelihood of structural damage or power outages.

In conclusion, the image of the red fire hydrant deep in the snow and the recently plowed sidewalk suggest that the city has faced a particularly severe winter season, with substantial snowfall that has likely caused various challenges and disruptions for its residents and infrastructure.

Table 11: One example used in in-context-learning to construct visual conversation data.

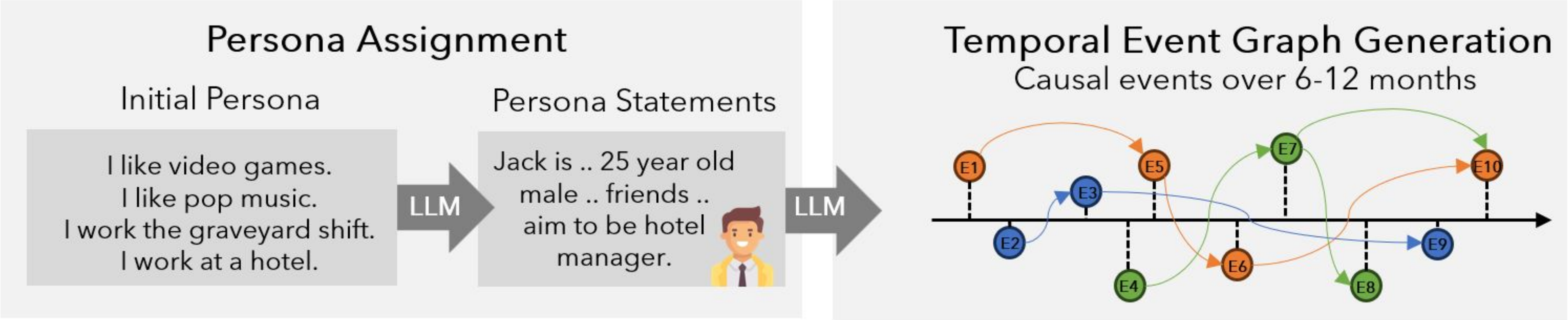
Step 1: Pre-define persona

Use few-liner personas from MSC dataset to create detailed personas

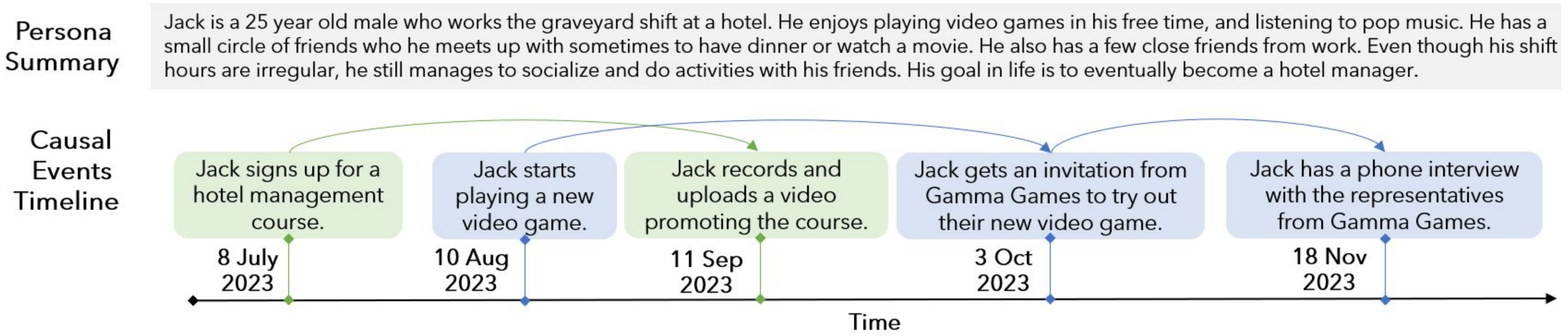


Step 2: Create event graphs

Use detailed persona to create event graphs

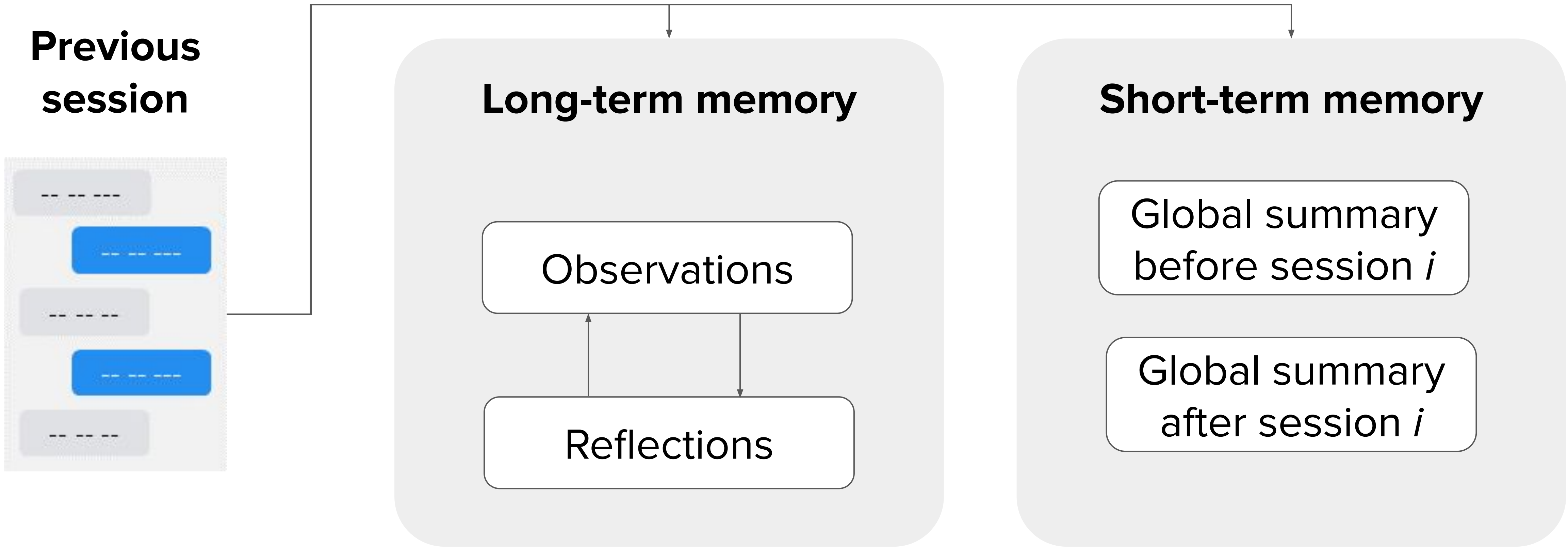


Persona & Event Graphs: Example



Prompts are designed to generate positive as well as negative events.
Events are generated in an iterative fashion.

Step 3: Memory system



Saved as a database and retrieved using encoder-decoder arch.

Always included in the prompt for generating conversations.

Observations: Example

Conversation (input)

1:56 pm, May 8, 2023

Hey Joanna! Long time no see! What's up? ..

Hey Nate! Long time no see! I've been working on a project lately - it's been pretty cool. What about you - any fun projects or hobbies?

I won my first video game tournament last week - so exciting!

Wow Nate! Congrats on winning! Tell me more - what game was it?

Thanks! it's a team shooter game.

Wow, great job! What was is called?

The game was called Counter-Strike: Global Offensive, and me and my team had a blast to the very end!

Sounds like a fun experience .. if I'm not into games.

...

Observations (output)

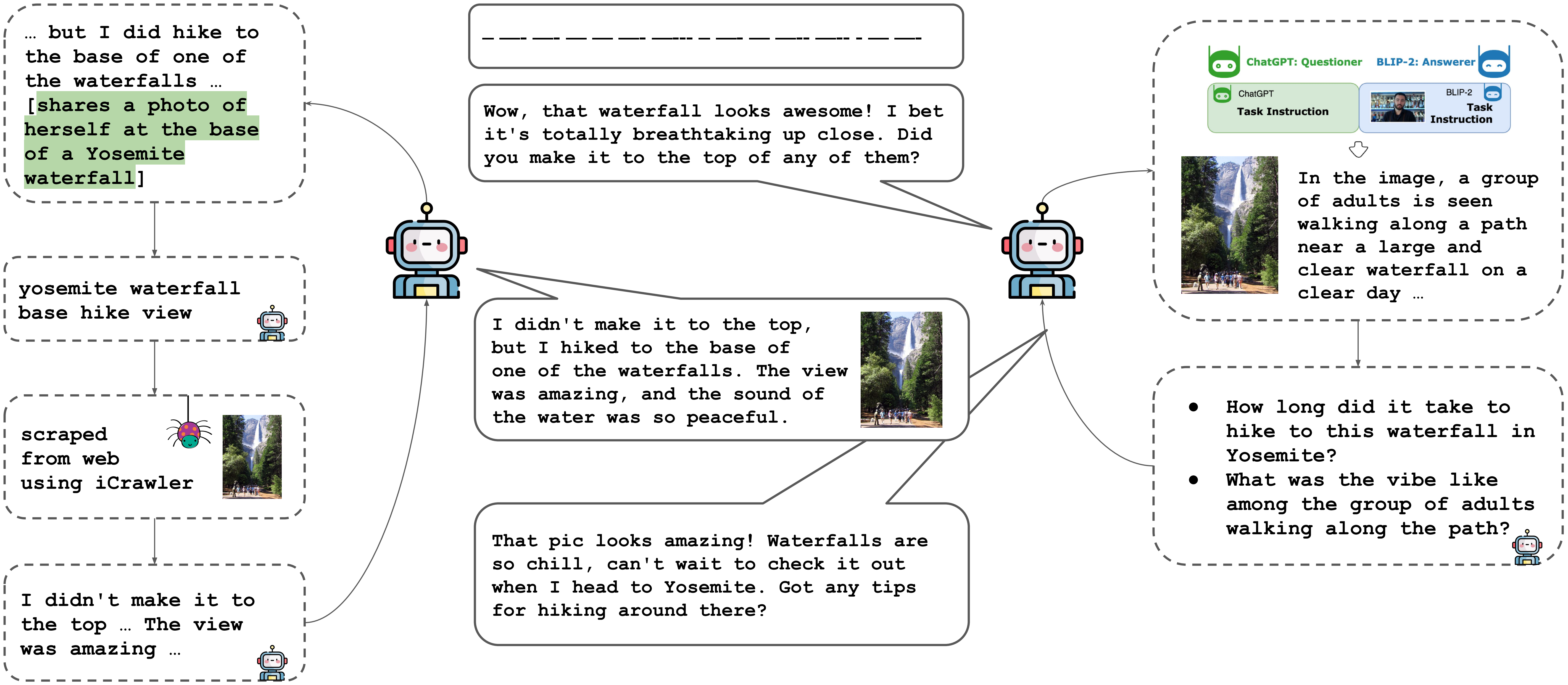
Joanna

- Joanna has been working on a project recently.
- Joanna enjoys writing, reading, watching movies, and exploring nature as hobbies.
- Joanna is into dramas and romcoms when it comes to movies.
- Joanna recommends a romantic drama movie that is all about memory and relationships.
- Joanna watched the recommended movie around 3 years ago and even owns a physical copy.

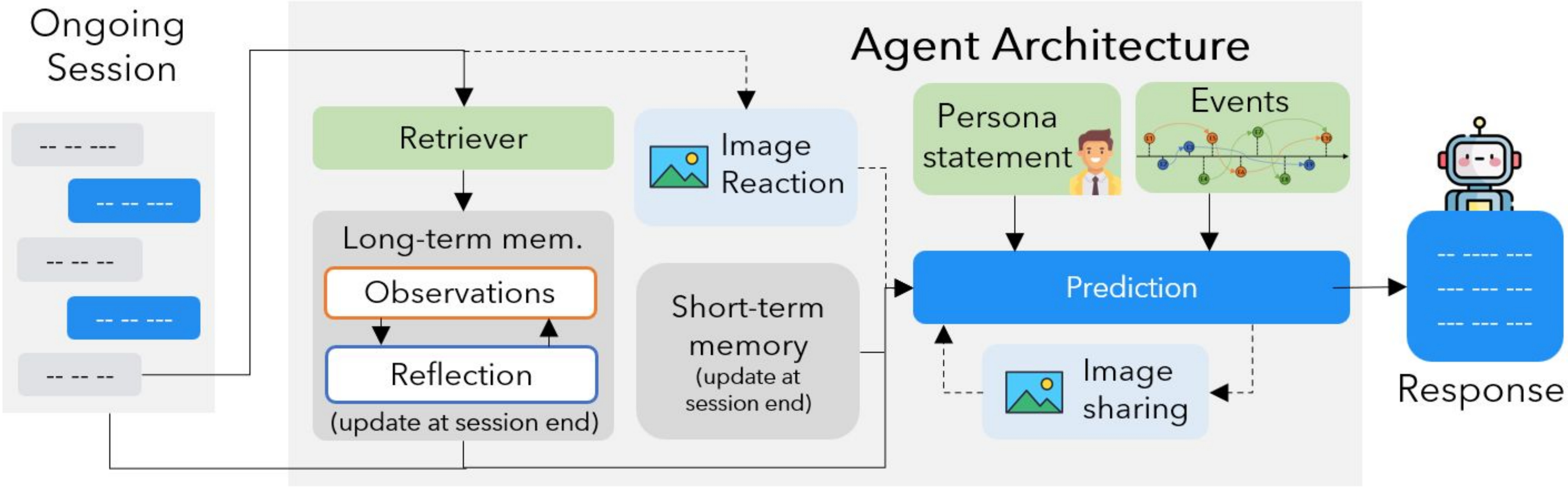
Nate

- Nate won his first video game tournament last week.
- The video game Nate won the tournament in is called Counter-Strike: Global Offensive.
- Playing video games and watching movies are Nate's main hobbies.
- Nate enjoys action and sci-fi movies.
- Nate loves watching classics.

Step 4: Multimodal Behaviour using LLMs

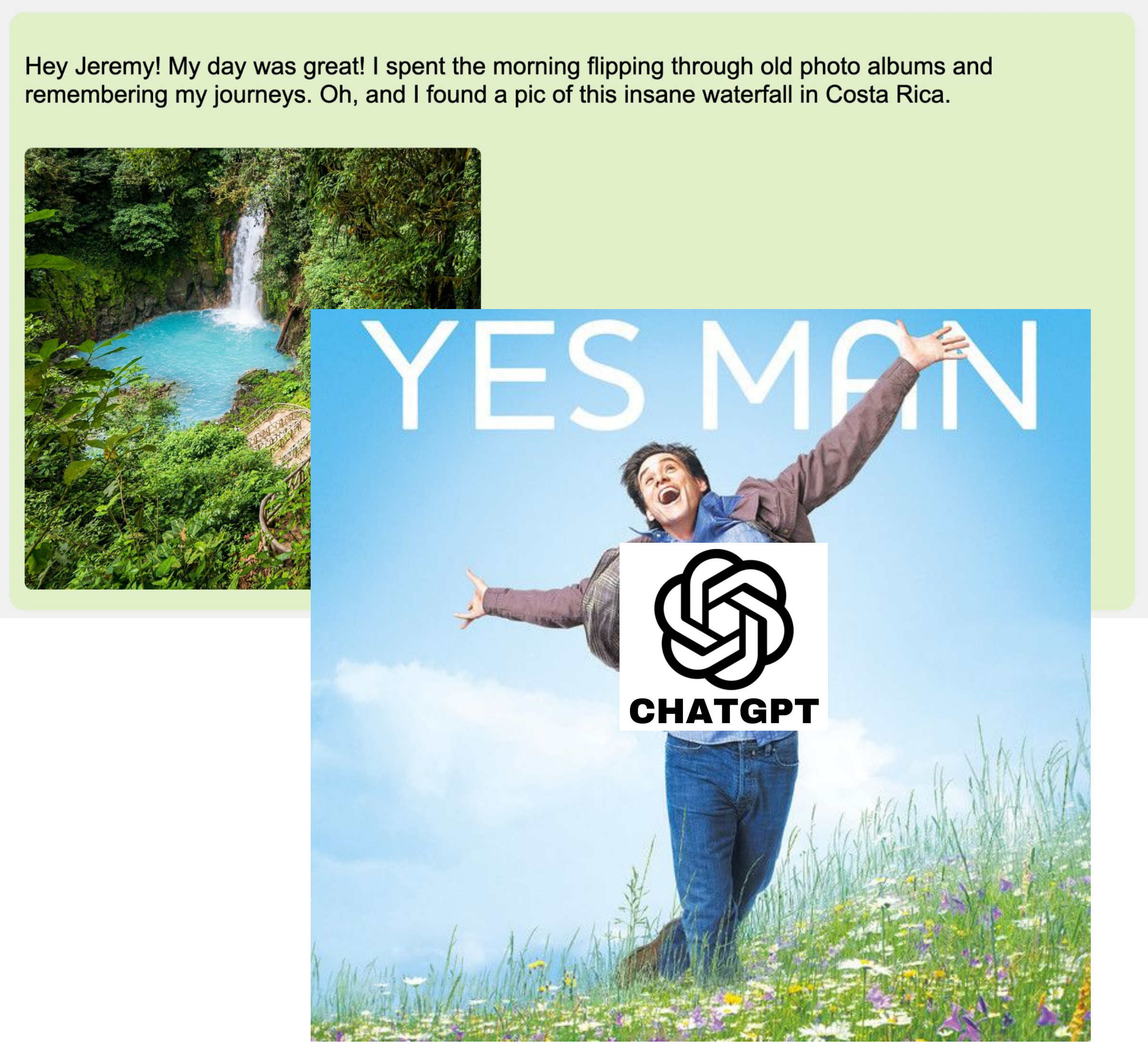


Bringing it all together



Persistent Issues

- Often, the images are generic and do not contain much multimodal information that we can ask questions about.
- Inconsistencies due to incorrectly retrieved information when generating next responses
- LLMs are bad at dialog repair ➔ they kinda 'go with the flow' ➔ skip events assigned to them
- On a similar note, LLMs are 'Yes Mans', they will say yes to most questions, at the cost of being inconsistent



Manual Editing

Remove or substitute irrelevant images

.. Oh, and I'm planning my solo trip to five countries! Exciting stuff.

Wow, Debra! 5 countries, awesome! Where're you headed? Need help planning or any tips? Count me in!

Edit inconsistent dialogs

My grandma sent me a postcard from Paris years ago.

.. Where'd you get it? …

Thanks Joe! I got the postcard from an antique shop…

Thanks Joe! My grandmother got the postcard from an antique shop…

Edit dialogs to follow event graph

Event: Joseph participates in a photography workshop and improves his photography skills.

.. Anything new at work or fun for the weekend?

.. I did a photoshoot last Friday and learned some new tricks. ..

.. I participated in a photography workshop last Friday and learned some new tricks. ..

LoCoMo Long-Term Conversational Dataset

Conversation Statistics	# Counts
Total. # conversations h.	50
Avg. # sessions k. in conversation h	19.3
Avg. # turns j. in session k	15.8
Avg. # tokens. conversation h	9,209.2
Avg. # tokens. dialogue h_{k_j} of turn j in session k	30.2
Avg. # tokens. observation o_{k_j} of turn j in session k	18.2
Avg. # tokens. summary w_k of session k	127.4

QA Annotations Guidelines: Example

(new) Questions that require aggregation of information (supported by edit scenarios 8a, 8b): These are the questions whose answer requires aggregation of information from different sessions. The information can be about anything, ranging from interests, personal life, names of people, professional life, habits, day-to-day life etc. about the speakers. For example,

D4:1	Hey Tiffany! What's been going on? Been super busy since we last talked. Just got back from France and it was awesome! Everything was stunning. Can't wait to catch up! 😊

D1:8	Cool! You liked the planetarium? I'm going to Japan next month - can't wait to explore the culture and check out the temples and shrines. Any trips you have coming up?

Question: Which countries has Joseph travelled to?
Answer: France, Japan
Evidence: D4:1, D1:8

Aggregation of information can also happen from images. For example,

D6:10 I like solving puzzles in my free time. Here's a pic of a tough one I just finished.

[shares a photo of a complex puzzle]

D8:5 That's awesome, Tiffany! Sounds like you have some cool stuff coming up. I've just finished a complex puzzle I've been working on for weeks - what a feeling! And I'm starting my postgrad classes next week, so I'm getting ready for that.

D8:9	Yeah, I've done crosswords and sudoku. They're tricky and addictive. Your finished crossword looks great, Tiffany. How long did you take to do it?

Question: What kind of puzzles does Joseph solve?
Answer: jigsaw puzzles, crosswords and sudoku
Evidence: D6:10, D8:5, D8:9

QA Task

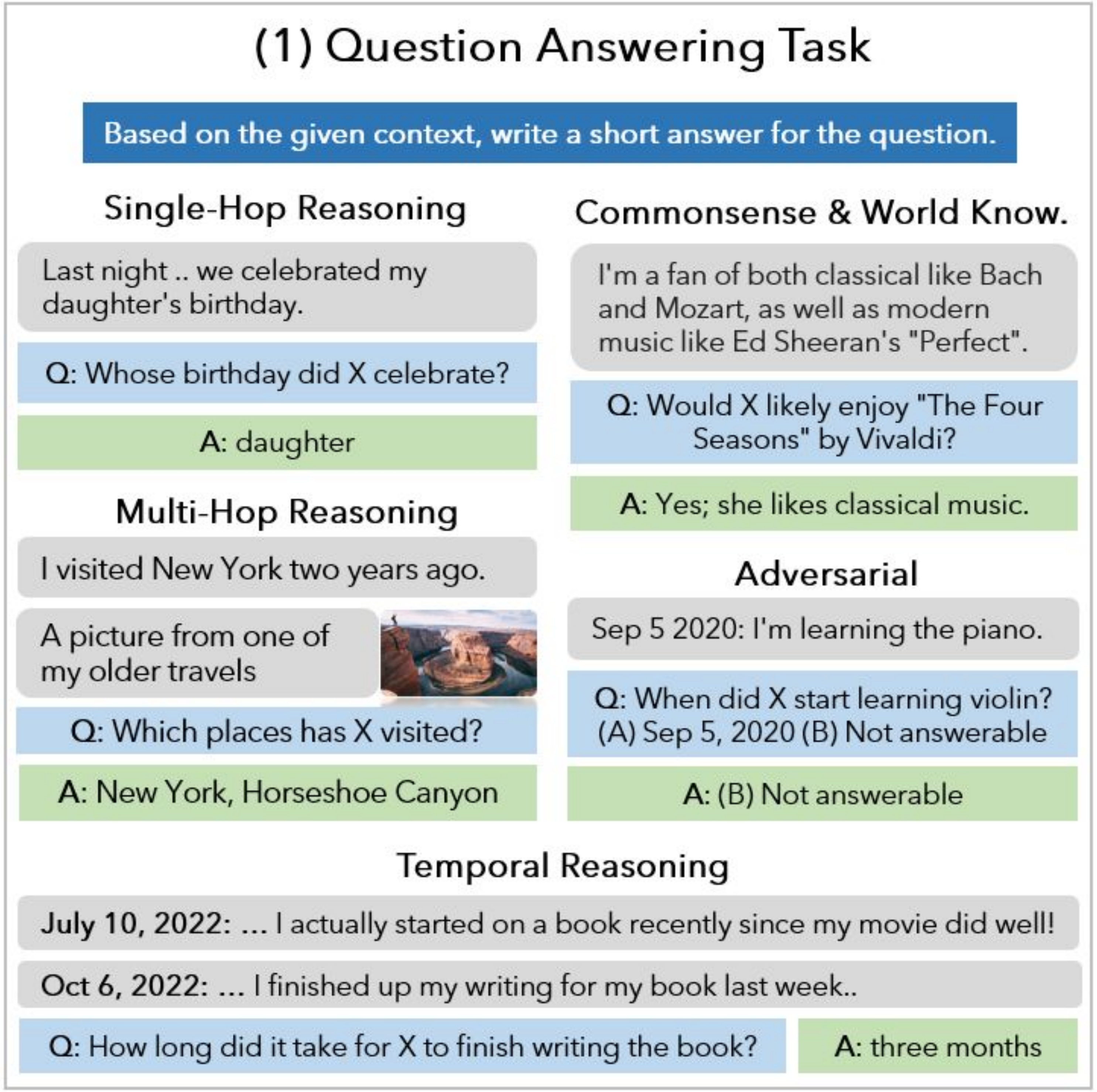
Five QA categories

Annotated:
- Multi-hop reasoning
- Commonsense & world knowledge reasoning
- Temporal reasoning

Generated:
- Single-hop reasoning
- Adversarial

Evaluation Metric: partial F1 score

QA Benchmark Statistics	
# questions. single-hop retrieval	2,705 (36%)
# questions. multi-hop retrieval	1,104 (14.6%)
# questions. temporal reasoning	1,547 (20.6%)
# questions. open domain knowledge	285 (3.9%)
# questions. adversarial	1,871 (24.9%)
Total. # questions.	**7,512**



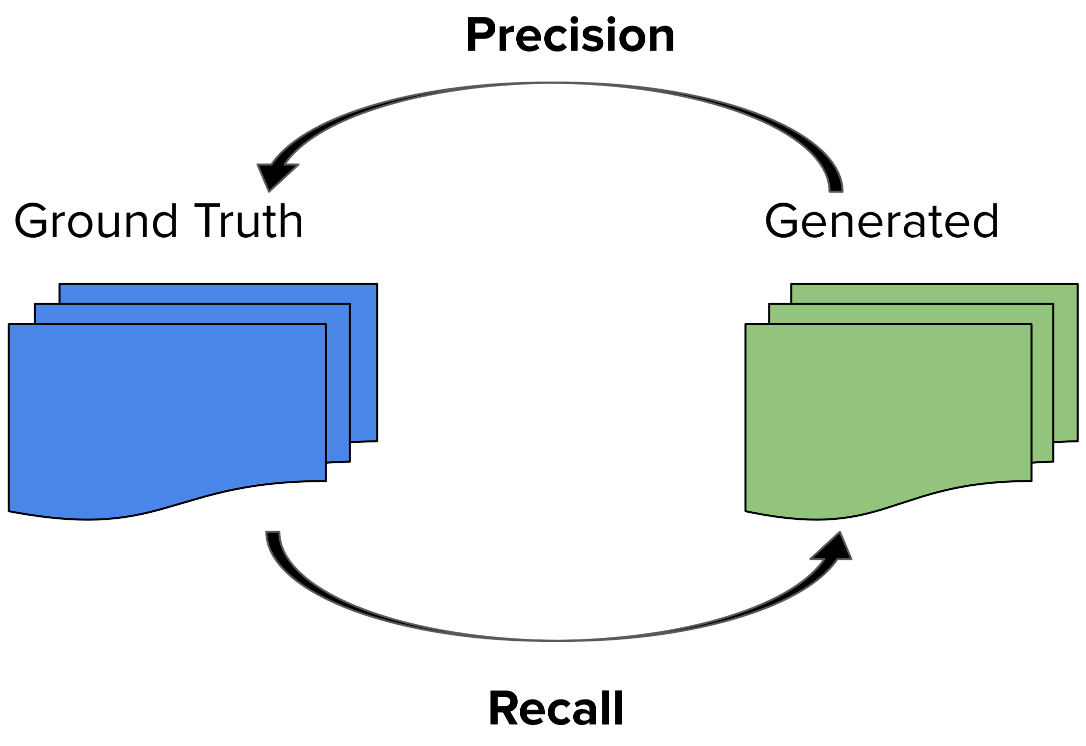
Event Summarization Task

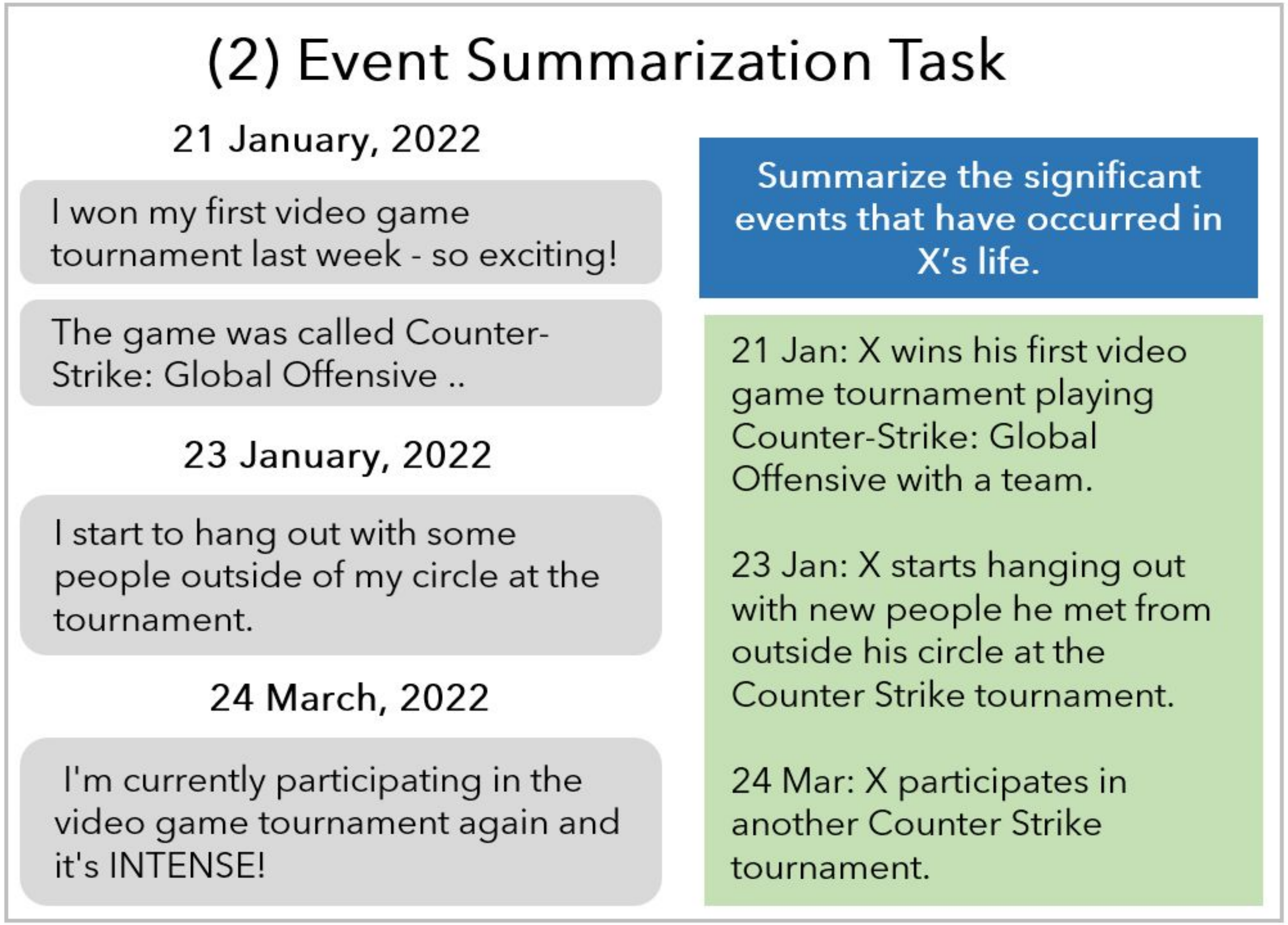
Event graphs are used as ground truth summaries.

Timestamps are the same as session dates.

Challenges include long-range temporal and causal reasoning.

Evaluation Metric: ROUGE, FactScore-F1





Multimodal Dialog Generation Task

Multi-modal Dialogue Generation Statistics	
Avg. # images. in conversation h	32.3

Challenges include retrieval of context and utilizing retrieved context correctly.

Evaluation Metric: MMRelevance, similar to CLIPScore

Experimental Setup

Baseline Models for QA

- *Long-context Models*
  - Gpt3.5-16k
- *Base LLMS*
  - GPT3.5
  - GPT4
  - Llama2-70B (instruction-tuned)
  - Mistral-7B (instruction-tuned)
- *Base LLMS + Retrieval*
  - Retriever: Dragon
  - Database type
    - Observation
    - Summary
    - raw dialog

Baseline Models for Summary

- *Long-context Models*
  - Gpt3.5-16k
- *Base LLMS*
  - GPT3.5
  - GPT4
  - Llama2-70B (instruction-tuned)
  - Mistral-7B (instruction-tuned)

Baseline Models for Dialog

- *Base MLLMS*
  - MiniGPT-5
- *Base MLLMs + Retrieval*
  - Retriever: Dragon
  - Database type:
    - Summary
    - observation

Results from QA Task

Category	Model	Context Length	Answer Prediction (F1)					
			Single Hop	Multi Hop	Temporal	Open Domain	Adversarial	**Overall**
Human	Human	-	95.1	85.8	92.6	75.4	89.4	87.9
Base	Mistral-Instruct-7B	8K	10.2	12.8	16.1	19.5	17.0	13.9
	Llama-2-Chat-70B	4,096	19.7	14.4	13.3	15.9	22.1	17.9
	GPT-3.5-turbo	4,096	**29.9**	23.3	**17.5**	**29.5**	12.8	22.4
	GPT-4-turbo	4,096	23.4	**23.4**	10.4	24.6	**70.2**	**32.1**
Long context	GPT-3.5-turbo-16K	4K	31.7	25.4	16.8	27.6	**13.1**	24.1
		8K	38.8	31.2	21.0	35.0	8.4	25.2
		12K	51.1	40.4	**25.0**	36.5	6.4	33.5
		16K	**56.4**	**42.0**	20.3	**37.2**	2.1	**37.8**

Table 2: **Question answering performance** of Base and Long-context models. Optimal performance is in **bold**. Results are based on F1-score for answer prediction; higher is better.

Results from QA Task with Retrieval

Retrieval Unit	top-k	Answer Prediction (F1 score) Single Hop	Multi Hop	Temporal	Open Domain	Adver- -sarial	**Overall**	Recall Accuracy (R@k) Single Hop	Multi Hop	Temporal	Open Domain	Adver- -sarial	**Overall**
None	-	29.9	23.3	17.5	29.5	12.8	22.4	-	-	-	-	-	-
Dialog	5	42.9	19.4	21.3	35.8	31.9	31.7	66.2	34.4	89.2	38.5	45.7	58.8
	10	46.3	26.8	24.8	37.5	**29.8**	34.6	72.8	247.4	97.3	53.8	54.3	67.5
	25	48.1	36.1	**26.2**	**43.4**	23.4	**35.8**	87.5	64.1	**97.3**	**67.9**	69.1	79.9
	50	**50.9**	**37.2**	24.6	38.3	17.0	34.8	**90.4**	**75.5**	97.3	67.9	**77.7**	**84.8**
Observation	5	44.3	30.6	41.9	40.2	**44.7**	**41.4**	52.9	40.1	81.1	38.5	29.8	49.6
	10	42.2	30.5	**42.1**	41.9	36.2	38.8	57.4	53.1	**83.8**	46.2	41.5	57.1
	25	**44.6**	33.2	41.8	41.9	27.7	38.0	**71.3**	63.8	83.8	66.7	45.7	66.0
	50	44.0	**34.5**	41.1	**41.9**	27.7	37.8	72.8	**73.2**	83.8	**74.4**	**56.4**	**71.1**
Summary	2	34.6	15.7	26.9	26.5	36.2	29.9	68.4	39.6	56.8	50.0	73.4	61.5
	5	**36.6**	**16.6**	**31.0**	**34.7**	38.3	**32.5**	81.6	57.0	70.3	60.3	86.2	75.1
	10	34.5	14.7	29.3	31.6	**40.4**	31.5	**93.4**	**82.3**	**91.9**	**80.8**	**94.7**	**90.7**

Table 3: **Question answering performance** of RAG-based GPT-3.5-turbo-16k. Optimal performance is in **bold**. Results are based on F1-score metric for answer prediction and recall@k for recall accuracy; higher is better.

Results from QA Task: Takeaways

- LLMs with limited context length face challenges in understanding extremely long conversations due to truncated context windows
- Long-context LLMs can comprehend longer narratives, yet they are prone to generating hallucinations.
- RAG is effective when conversations are stored as observations
- **Time reasoning** and **open-domain knowledge** questions are the most challenging scenarios

LLMs struggle with change of state over time

Session 2

Tiffany reads her first sci-fi book, “The Time-Traveller’s Daughter”

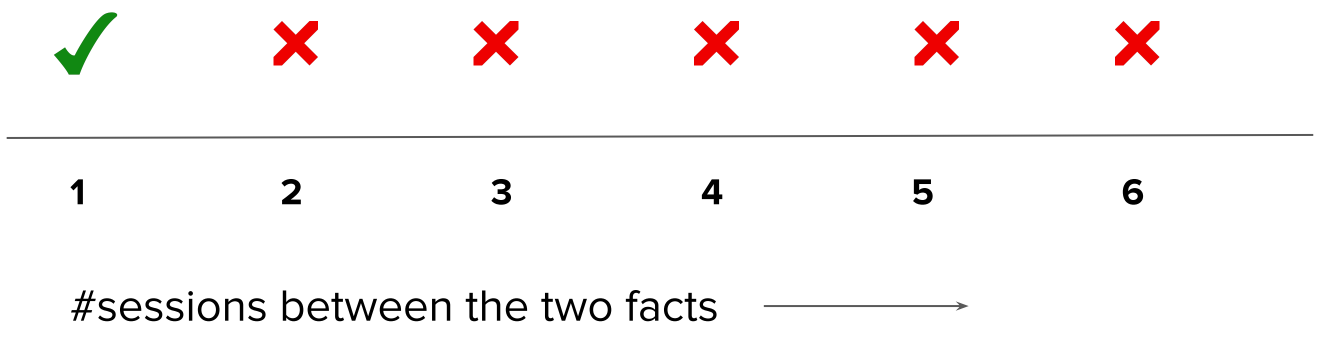
Session 10

Tiffany reads another sci-fi book, “The Maze of Time”

Question: What is the first sci-fi book Tiffany has read?

Correct Answer: The Time-Traveller’s Daughter

GPT3.5 [16k]: The Maze of Time



LLMs struggle with attribution

Speaker A

Collects vintage coins

Speaker B

Collects vintage coins, stamps, maps, postcards

Speaker B

Has a favorite vintage postcard from Paris sent by grandmother

Question: What is Speaker A's favorite postcard?
Correct Answer: There is no information about Speaker A's favorite postcard in the conversation.
GPT3.5 [16k]: Speaker A's favorite postcard is a picture of a vintage coin from Paris.

Question: What is Speaker B's memory about their grandmother?
Correct Answer: Speaker B's grandmother sent them a vintage postcard from Paris.
GPT3.5 [16k]: A cherished memory for Speaker B about their grandmother is receiving a vintage coin from her as a gift.

LLMs struggle with coreference resolution over long context

D2:8 That's awesome! I had no idea quantum computing could do that. I found a book called **"The Time-Traveller's Daughter"** which looks interesting. It's my first foray into sci-fi. I am reading a few pages before going to bed most nights. Do you have anything fun coming up?

D10:4 That sounds great, Joseph! I joined a book club recently and introduced them to the book I was reading, they loved it. We also read this sci-fi novel "The Maze of Time". It was interesting and made me think. Have you ever read any sci-fi books that make you question reality?

Question: Which books has Tiffany's book club read?

Correct Answer: The Time Traveller's daughter, The Maze of Time, Three-Body Problem

Evidence: D2:8, D10:4, D10:16

GPT3.5 [16k]: The Maze of Time, Three-Body Problem, Dragon Tattoo.

When do models struggle to retrieve context?

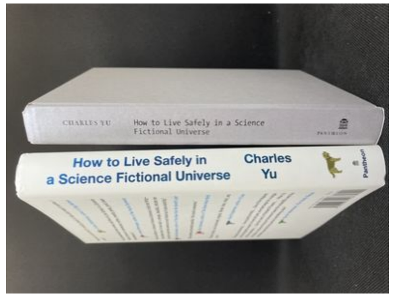
- No explicit overlap between question and answer

Which science topic did Joseph and Tiffany talk about?

Quantum computing

- Images needing some amount of detailed understanding not covered in caption

Which book does Joseph share with Tiffany?



- Dialog context is scattered over many adjacent dialogs, such as question followed by answer

How does Tiffany like to stay motivated?

By using a vision board

Results from Summarization Task

Category	Model	Context Length	ROGUE			FactScore		
			ROGUE-1	ROGUE-2	ROGUE-L	Precision	Recall	F1
Base	Mistral-Instruct-7B	8K	29.4	7.2	14.1	27.1	19.8	23.0
	Llama-2-Chat-70B	4,096	28.1	9.3	14.8	36.3	22.7	28.3
	GPT-4-turbo	4,096	38.8	11.4	20.6	**51.6**	41.8	45.1
	GPT-3.5-turbo	4,096	**41.1**	**13.5**	**20.9**	45.3	**46.5**	**45.9**
Long context	GPT-3.5-turbo-16K	16K	36.2	8.5	16.4	42.3	37.8	39.9

Table 4: **Event summarization performance** of Base and Long-context models. The optimal performance is shown in **bold**. Results are based on ROUGE and FactScore (Min et al., 2023) metrics; higher is better.

- The use of incremental summarization with gpt-3.5-turbo leads to the highest performance in both recall and F1 score.
- The long-context model does not surpass the base model.

Results from Summarization Task: Error Types

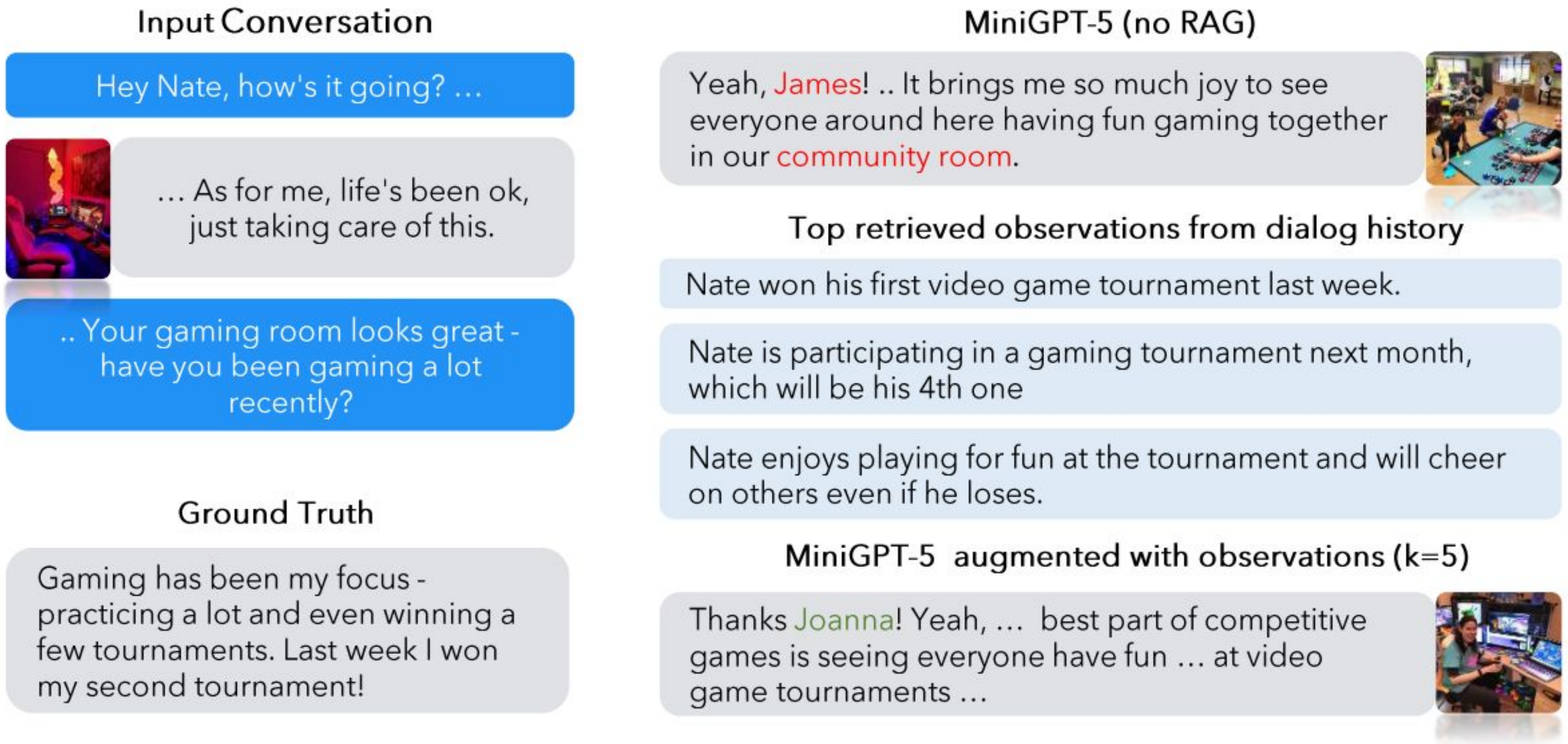
Error Type	Explanation	Ground truth event *or* relevant dialogs	Predicted event
Missing information	Key details about event are omitted because the model fails to make causal and temporal connections over a long conversation.	Joanna submits her third screenplay on loss, identity, and connection to a film contest	Joanna submits her recent screenplay to a film contest.
Hallucination	Non-existent details or details from a different event are padded onto an event	*N*: 'The gaming party was a great success!' *N*: '... said they'd want to do it again next month!' *N*: 'On another note, I made vegan ice cream ...'	Nate's vegan ice cream is a huge success and people want to do it again next month.
Misunder- -standing of dialog cues	e.g., model confuses a light-hearted statement from a speaker as a serious statement	*J*: '.. these trails that made me feel like writing a drama.' *N*: '.. go together .. Maybe I'll start to think of a drama myself and write a screenplay ...' *J*: 'Haha, now that would be something! ...'	Nate considers writing his own drama screenplay.
Speaker attribution	Event is attributed to the wrong speaker	Nate invites Joanna to try his homemade lactose-free ice cream.	Joanna invites Nate to her home to try her dairy-free ice cream recipe.
Saliency	Unimportant interactions in the conversation are considered significant by model	*N*: Hey Joanna, what's been up since we last chatted? How's it going?	Nate asks Joanna how she has been she they last talked.

Table 7: **Taxonomy of errors in LLM-generated event summaries.** Five types of errors predominantly occur in the event summaries generated by LLMs. Examples are based on predictions from `gpt-3.5-turbo`.

Results from Multimodal Dialog Generation Task

Category	top-k	BLEU-1/2	Rouge-L	MM-R
Base	-	57.1 / 34.2	12.4	56.1
+ summary	1	58.2 / 34.1	12.8	56.9
+ summary	2	56.5 / 32.8	12.1	55.1
+ summary	5	56.1 / 32.5	12.0	55.2
+ observation	5	**59.7 / 35.1**	**13.6**	**57.8**
+ observation	10	59.1 / 34.9	12.8	57.1
+ observation	25	58.5 / 34.2	12.0	56.5

Table 6: **Multi-modal dialogue generation performance** comparison between different training variants of MiniGPT-5. The optimal performance is shown in **bold**.



A. Example of a prediction from MiniGPT-5 with and without retrieval-based augmentation

- Incorporating context into training enhances performance, with the inclusion of observation as context yielding significantly improved results.

Prevailing challenge: How to create data for multimodal change of state?

What we want:

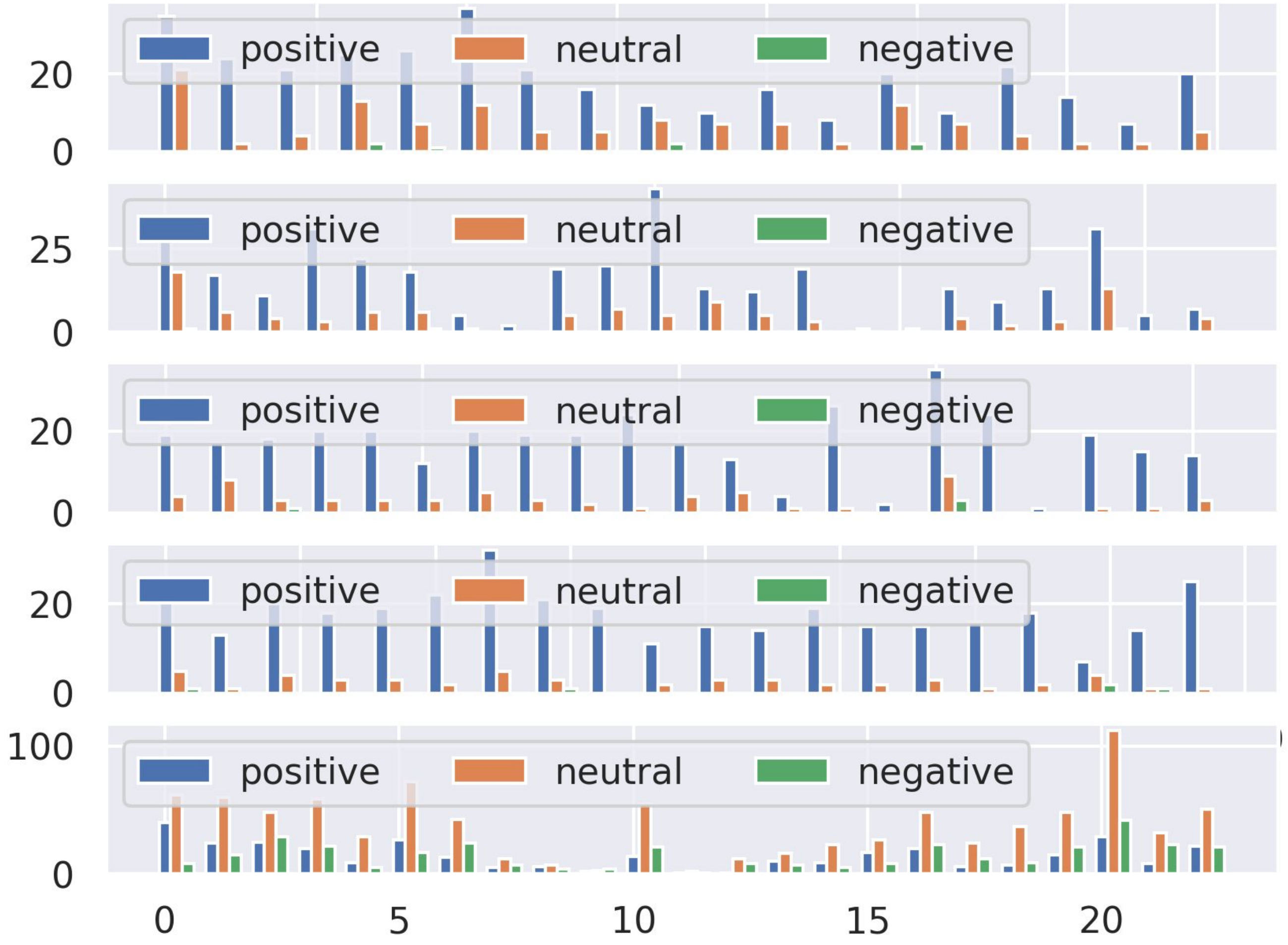
Change of state

Multiple values of same attribute

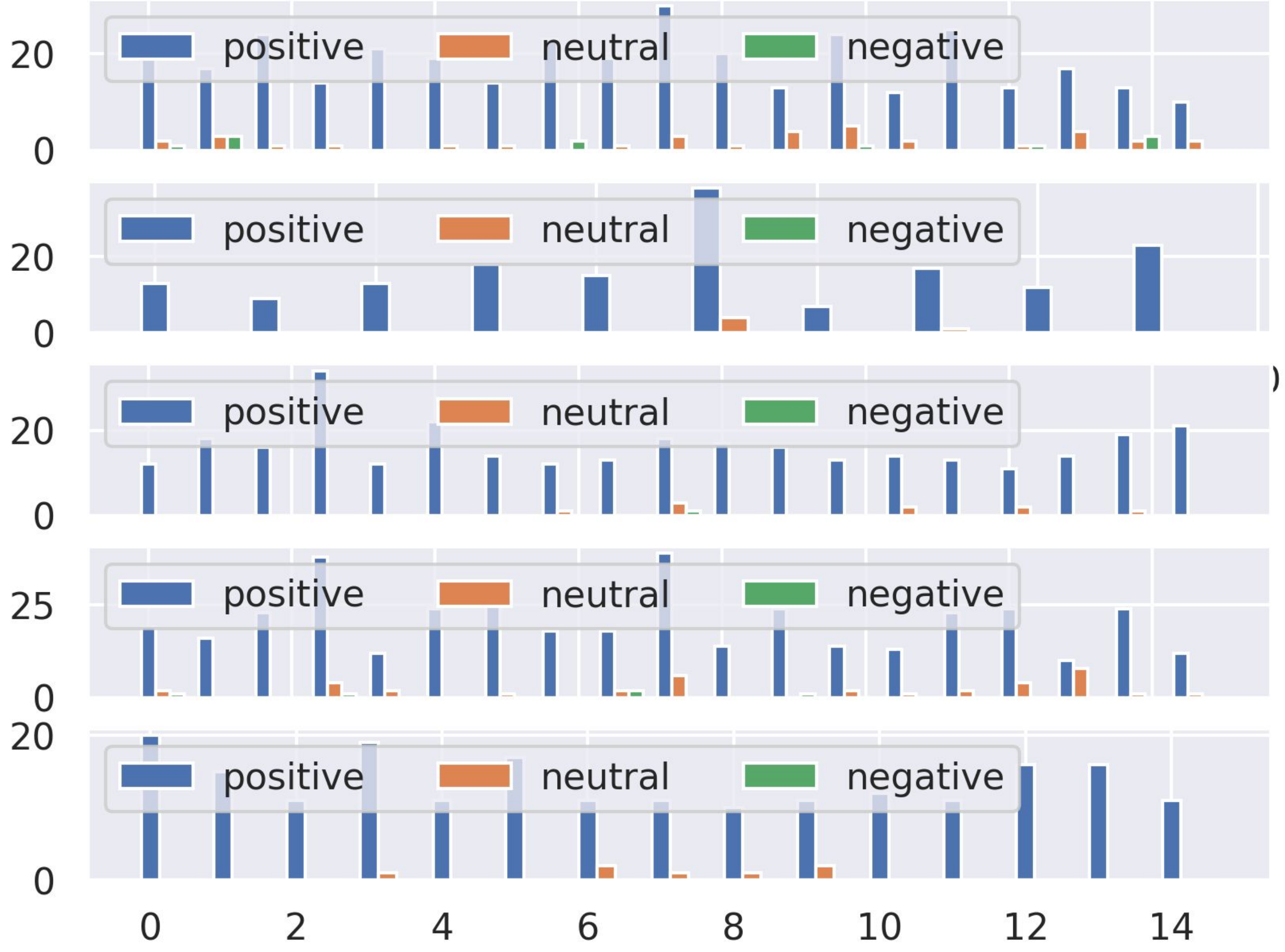
What we have: Standalone images

Prevailing challenge: Real-world vs. semi-synthetic conversations?

Variation in sentiment over time in real-world conversations



Variation in sentiment over time in synthetic conversations



Thank you!