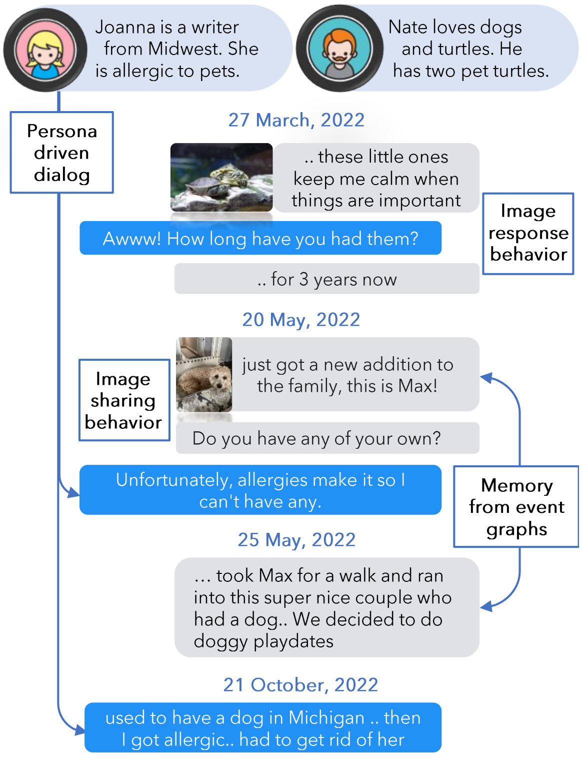
Evaluating Very Long-Term Conversational Memory of LLM Agents

Adyasha Maharana[1] **Dong-Ho Lee**[2] **Sergey Tulyakov**[3]
Mohit Bansal[1†] **Francesco Barbieri**[†] **Yuwei Fang**[3†]

University of North Carolina, Chapel Hill[1] University of Southern California[2] Snap Inc.[3]

Abstract

Existing works on long-term open-domain dialogues focus on evaluating model responses within contexts spanning no more than five chat sessions. Despite advancements in long-context large language models (LLMs) and retrieval augmented generation (RAG) techniques, their efficacy in *very* long-term dialogues remains unexplored. To address this research gap, we introduce a machine-human pipeline to generate high-quality, *very* long-term dialogues by leveraging LLM-based agent architectures and grounding their dialogues on personas and temporal event graphs. Moreover, we equip each agent with the capability of sharing and reacting to images. The generated conversations are verified and edited by human annotators for long-range consistency and grounding to the event graphs. Using this pipeline, we collect LoCoMo, a dataset of *very* long-term conversations, each encompassing approx. 600 turns and 16K tokens on avg., over up to 32 sessions. Based on LoCoMo, we present a comprehensive evaluation benchmark to measure long-term memory in models, encompassing question answering, event summarization, and multi-modal dialogue generation tasks. Our experimental results indicate that LLMs exhibit challenges in understanding lengthy conversations and comprehending long-range temporal and causal dynamics within dialogues. Employing strategies like long-context LLMs or RAG can offer improvements but these models still substantially lag behind human performance.[1]



Joanna is a writer from Midwest. She is allergic to pets.

Nate loves dogs and turtles. He has two pet turtles.

.. these little ones keep me calm when things are important

Awww! How long have you had them?

.. for 3 years now

just got a new addition to the family, this is Max!

Do you have any of your own?

Unfortunately, allergies make it so I can't have any.

... took Max for a walk and ran into this super nice couple who had a dog.. We decided to do doggy playdates

used to have a dog in Michigan .. then I got allergic.. had to get rid of her

Figure 1: **An example in LoCoMo.** Dialogs are steered by the speakers' personas and corresponding events e.g., Joanna's responses are consistent with her pet allergies. For Nate, the event *got a new dog* is followed by a *playdate with neighbor's dog*, showcasing long-term memory. Multimodal dialog is enabled with image-sharing and image-response behaviors.

1 Introduction

Despite recent advancements in dialogue models based on LLMs for extended contexts (Bertsch et al., 2023; Xiao et al., 2023), as well as the integration of retrieval augmented generation (RAG) techniques (Shuster et al., 2021; Ram et al., 2023; Shi et al., 2023), there is still a need for thorough evaluation of their efficacy in handling very long conversations. Indeed, studies in long-term open-domain dialogues have concentrated on assessing model responses within limited contexts e.g., ~1K tokens over five chat sessions (Xu et al., 2022; Jang et al., 2023b; Zhang et al., 2023). This long term evaluation is crucial for refining engaging chatbots capable of remembering key information from past interactions, to generate empathetic, consistent, and useful responses.

---

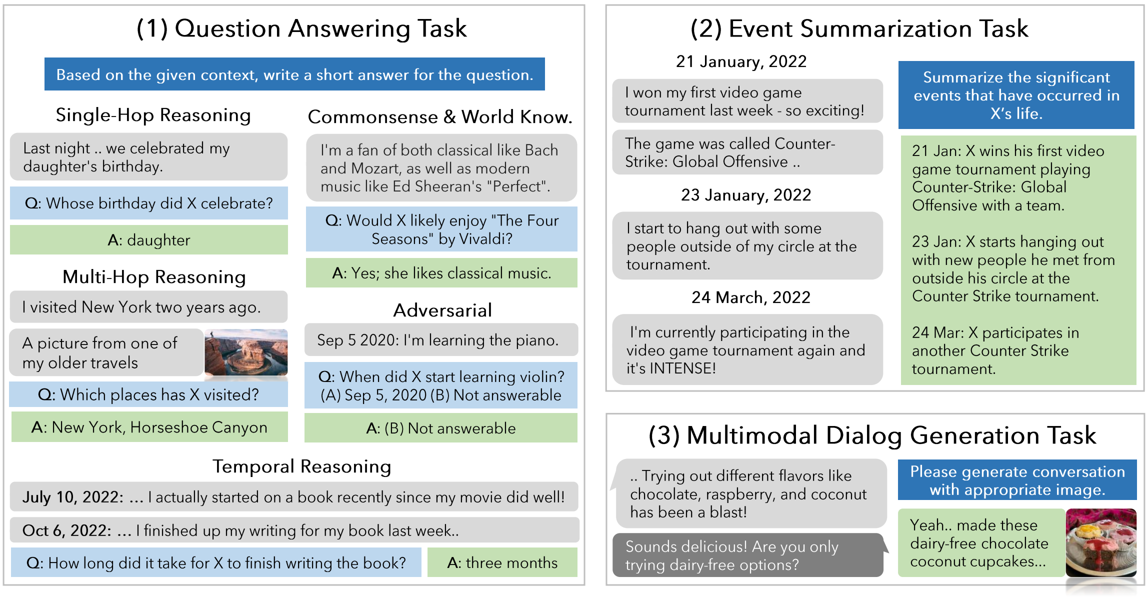
[1] Code and data are at `https://snap-research.github.io/locomo`

[†] Equal advising.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13851–13870
August 11-16, 2024 ©2024 Association for Computational Linguistics

Dataset	Avg. turns per conv.	Avg. sessions per conv.	Avg. tokens per conv.	Time Interval	Multimodal	Collection
MPChat (Ahn et al., 2023)	2.8	1	53.3	-	✓	Reddit
MMDialog (Feng et al., 2022)	4.6	1	72.5	-	✓	Social media
Daily Dialog (Li et al., 2017)	7.9	1	114.7	-	✗	Crowdsourcing
SODA (Kim et al., 2023)	7.6	1	122.4	-	✗	LLM-generated
MSC (Xu et al., 2022) (train; 1-4 sessions)	53.3	4	1,225.9	few days	✗	Crowdsourcing
Conversation Chronicles (Jang et al., 2023a)	58.5	5	1,054.7	few hours - years	✗	LLM-generated
LoCoMo (ours)	588.2	27.2	16,618.1	few months	✓	LLM-gen. + crowdsourc.

Table 1: **Statistics of LoCoMo** compared to existing dialog datasets. The average length of a conversation in LoCoMo is 16x that of MSC (Xu et al., 2022), distributed over 10x more turns and 5x more sessions (on average).


(1) Question Answering Task
Single-Hop Reasoning
Commonsense & World Know.
Multi-Hop Reasoning
Adversarial
Temporal Reasoning
(2) Event Summarization Task
(3) Multimodal Dialog Generation Task

Based on the given context, write a short answer for the question.

Last night .. we celebrated my daughter's birthday.

Q: Whose birthday did X celebrate?

A: daughter

I visited New York two years ago.

A picture from one of my older travels

Q: Which places has X visited?

A: New York, Horseshoe Canyon

I'm a fan of both classical like Bach and Mozart, as well as modern music like Ed Sheeran's "Perfect".

Q: Would X likely enjoy "The Four Seasons" by Vivaldi?

A: Yes; she likes classical music.

Sep 5 2020: I'm learning the piano.

Q: When did X start learning violin? (A) Sep 5, 2020 (B) Not answerable

A: (B) Not answerable

July 10, 2022: ... I actually started on a book recently since my movie did well!

Oct 6, 2022: ... I finished up my writing for my book last week..

Q: How long did it take for X to finish writing the book?

A: three months

21 January, 2022

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

The game was called Counter-Strike: Global Offensive ..

23 January, 2022

I start to hang out with some people outside of my circle at the tournament.

24 March, 2022

I'm currently participating in the video game tournament again and it's INTENSE!

Summarize the significant events that have occurred in X's life.

21 Jan: X wins his first video game tournament playing Counter-Strike: Global Offensive with a team.

23 Jan: X starts hanging out with new people he met from outside his circle at the Counter Strike tournament.

24 Mar: X participates in another Counter Strike tournament.

.. Trying out different flavors like chocolate, raspberry, and coconut has been a blast!

Sounds delicious! Are you only trying dairy-free options?

Please generate conversation with appropriate image.

Yeah.. made these dairy-free chocolate coconut cupcakes...

Figure 2: **Overview of our evaluation framework**. We propose three tasks: question answering, event summarization and multimodal dialog generation to evaluate models' comprehension in very long-term dialogues.

To this end, we present the first study of very long-term open-domain multi-modal dialogues, closely mirroring real-world online interactions, collected via a human-machine pipeline where we first use LLM-based generative agents to generate conversations and then ask human annotators to fix any long-term inconsistencies in the conversations. Specifically, drawing on the understanding that real-world conversations are a complex blend of collective memories (Assmann and Czaplicka, 1995; Hirst and Manier, 2008), individual viewpoints (Hirst et al., 2018), external influences (Hirst and Echterhoff, 2012), and the unique persona of the speakers (Pruitt and Grudin, 2003; Cooper, 1999; Zhou et al., 2020; Shum et al., 2019), we create *very long-term* dialogues based on LLM agent with the following features: (1) a unique persona (§3.1); (2) a timeline of causally interlinked events in their lives (§3.2); and (3) *reflect & response* mechanism to respond based on dialogue history (like in Park et al. (2023)) and *image sharing & image reaction* behavior which sends or reacts to images (§3.3). Finally, human annotators fix long-range inconsistencies in dialogues, remove irrelevant images, and verify the grounding of dialogs to events (§3.4). With this pipeline, we create LoCoMo, a dataset of 10 *very long-term* dialogues, each consisting of 600 turns and 16K tokens on avg., over up to 32 sessions (see Figure 1, Table 1).

Conventional approaches for evaluating conversational agents in open-domain dialogues involves directly evaluating the agent response based on past dialogue history. It often employs lexical overlap (Papineni et al., 2002) and semantic overlap (Zhang et al., 2019) between ground truth and the agent response, or consistency (Ghazarian et al., 2022), contradiction (Nie et al., 2021; Welleck et al., 2019), and empathy (Zhang et al., 2021a, 2022) of the agent response. However, these evaluation metrics are not well-suited for directly assessing the agent's comprehension of long-term contexts. In this study, we present a holistic evaluation framework to assess an agent's proficiency in managing and responding within long-term contexts (see Figure 2). First, agents need to "recall" past context correctly to integrate relevant information into future responses. We present a direct examination of their *memory* via a *question answering* (QA) task (§4.1). We classify questions into five distinct reasoning types to evaluate memory from multiple perspectives: single-hop, multi-hop, temporal, commonsense or world knowledge, and adversarial. Second, agents also need to recognize long-range causal and temporal connections in the dialogues to generate empathetic and relevant responses. We propose a measurement of their causal and temporal understanding with an *event graph summarization* task (§4.2). In this task, the event graphs linked to each LLM speaker serve as the correct answers, and models are tasked with extracting this information from the conversation history. Third, conversational agents need to utilize relevant context recalled from past conversations to generate responses that are consistent with the ongoing narrative. We assess this ability via the *multi-modal dialog generation* task (§4.3).

We present extensive experimental results on the LoCoMo benchmark using instruction-based LLMs, long-context LLMs, and RAG techniques (§5). Our findings include: (1) Long-context LLMs and RAG demonstrate effectiveness in QA tasks, improving 'memory' capabilities of LLMs (with improvements ranging from 12-20%), but still significantly lag behind human levels (by 36%), especially in temporal reasoning, (by 41%); (2) long-context LLMs demonstrate significant difficulty with adversarial questions in the QA task, showing a performance that is 65% lower than the base model. They are especially prone to misassigning dialogs or events to the wrong speaker. Moreover, they show poor performance on *event graph summarization*, indicating that they may grasp the factual elements within the entire conversation but do not accurately comprehend the context; and (3) RAG offers a balanced compromise, combining the accuracy of short-context LLMs with the extensive comprehension of wide-context LLMs, and does particularly well when dialogues are transformed into a database of assertions (*observations*) about each speaker's life and persona.

2 Related Work

Long-term Dialogue. Recent approaches involve retrieving historical context from a range of previous dialogues and reasoning over retrieved segments in a temporal order (Lee et al., 2023b; Lu et al., 2023; Zhong et al., 2023; Liang et al., 2023) and/or using events to scaffold the dialogues (Jang et al., 2023b; Zhang et al., 2023) to enable consistency in long-term conversations. Some limitations of such frameworks are: (1) The accuracy of retrieval can be compromised, as the retrieval model is generally trained on tasks focusing on semantic similarity rather than specifically on such dialogues. Additionally, real-world dialogues often feature co-references and missing content (*i.e.,* anaphora) (Anantha et al., 2021), which further complicate the retrieval process (Mallen et al., 2023; Gao et al., 2023b; Liu et al., 2023b); (2) Challenges arise in reasoning over retrieved documents, especially when the model struggles to identify the correct context among the retrieved data (Liu et al., 2023a); (3) Reasoning over time intervals presents challenges. For example, the way a system responds about past events can vary depending on the amount of time that has passed since the last conversation (Zhang et al., 2023; Jang et al., 2023b). Therefore, it is essential to have conversations of considerable length, as well as a systematic evaluation framework, to accurately assess the effectiveness of approaches to long-term dialogue generation. We design a long-term conversation generation pipeline based on retrieval augmentation and events graphs and propose a framework for evaluating long-term dialog agents.

Multi-modal Dialogue. Multi-modal dialogue primarily consists of two types of tasks: image-grounded dialogue and image-sharing dialogue. The image-grounded dialogue task is centered around responding to questions (Antol et al., 2015; Das et al., 2017; Kottur et al., 2019) or creating natural conversations related to specific images (Mostafazadeh et al., 2017; Shuster et al., 2018; Meng et al., 2020; Zheng et al., 2021). Conversely, the image-sharing dialogue task focuses on selecting images that semantically align with the provided dialogue context (Zang et al., 2021; Feng et al., 2022; Lee et al., 2023c). We use a method from the image-sharing dialogue task to create multimodal dialogs which are then evaluated as an image-grounded dialogue task.

Synthetic Evaluation Benchmark. Faced with a shortage of human-generated data and observing that LLMs are approaching the quality of human-level annotations (He et al., 2023; Lee et al., 2023a), there has been a surge in research drawing inspiration from this development. Consequently, numerous studies have started utilizing LLMs to augment or synthesize large-scale dialogue benchmarks for assessing responses in everyday social interactions (Kim et al., 2023), examining responses in multi-modal environment (Feng et al., 2022), and evaluating responses that align with specific persona (Jandaghi et al., 2023). We leverage LLMs to create data but ensure its high quality with human verification and editing.

3 Generative Pipeline for LOCOMO

An overview of our generative pipeline for LOCOMO is shown in Figure 3. We create two virtual agents, named $\mathcal{L}_1$ and $\mathcal{L}_2$, each initialized with a LLM $\mathcal{M}$ (*i.e.,* `gpt-3.5-turbo`). To start, unique persona statements p are assigned to each agent $\mathcal{L}_i$, ensuring the integration of distinct personalities into their dialogues (§3.1). To mirror real-life experiences, we create a temporal event graph $\mathcal{G}$ for each agent, which illustrates a realistic sequence of life events (§3.2). The LLM agent architecture (Park et al., 2023) is utilized for each agent $\mathcal{L}_i$, enabling them to effectively memorize and reflect conversation history into ongoing dialogues (§3.3). Further, each agent $\mathcal{L}_i$ can share coherent images, thereby enhancing the multi-modal dialogue aspect. Finally, human annotators are tasked with manually filtering and refining the generated data (§3.4).

3.1 Persona

We select an initial persona statement p_c from the MSC dataset (Xu et al., 2022), encompassing 4 to 5 sentences, and employ `gpt-3.5-turbo` as $\mathcal{M}$ to expand these into full persona statement p (See examples and prompt details in Appendix A.1). The generated statements typically include details about one or more of the following elements (Gao et al., 2023a): objectives, past experiences, daily habits, and interpersonal relationships, as well as name, age, and gender of the individual.

3.2 Temporal Event Graph

To utilize the real-life experiences of each agent in the conversation, we construct a temporal event graph, labeled as $\mathcal{G}$, for each agent. This graph $\mathcal{G}$, consisting of events e_i, is produced by applying the condition of $\mathcal{M}$ (i.e., `text-davinci-003`) on a designated persona p. Each event e_i is associated with a date of occurrence t_i. $\mathcal{G}$ includes causal connections $l = (e_i, e_j)$ that illustrate the causal relationships among events $e_i \in \mathcal{G}$ and reflect a natural succession of events in an individual's life. For each $\mathcal{G}$, we create up to 25 events, spread across a time frame of 6 to 12 months, in an iterative process that balances between inference time and the coherence of temporal and causal connections in the timeline. Initially, a small batch of $k = 3$ events is generated, which is then used iteratively as input prompt to create the subsequent batch of k events. See details in Appendix A.2.

3.3 Virtual Agent Architecture

Every agent $\mathcal{L}_i$ incorporates modules from generative agent architecture (Park et al., 2023). The agent has two functions: (1) *reflect & respond*; and (2) *image sharing & image reaction*. The agent is asked to primarily use the *reflect & respond* function while employing *image sharing & image reaction* function judiciously and appropriately within the context of the conversation.

Reflect & Respond. The fundamental process for each agent to *reflect and respond* involves the concept of short-term and long-term memory. During inference, agent $\mathcal{L}_i$ conditions its responses on both short and long-term memories, paralleling how humans remember recent conversations while also recalling distilled important experiences from long-term memory. Following each session k, each agent is asked to produce a summary w_k that is then stored in the short-term $\mathcal{H}_s$. This summary w_k is generated by conditioning $\mathcal{M}$ on both the most recent session conversation history h_k and the preceding summary $w_{k-1} \in \mathcal{H}_l$. For each turn j within session k, a single turn of the conversation h_{k_j} is transformed into an observation o_{k_j} and then stored in the long-term memory $\mathcal{H}_l$. Then, agent $\mathcal{L}_i$ generates a response in session $k+1$ on the date t^s_{k+1} by basing it on the latest summary w_k, reflections based on the retrieved relevant observations $o \in \mathcal{H}_s$, the ongoing conversation history in the current session h_{k+1} and persona statement p. Long-term temporal narratives are induced in the conversation by additionally conditioning the agent's response on the subset of events in $\mathcal{G}$ that occur between the last and current session i.e. $\{e \in \mathcal{G} \,|\, t^s_k \, < \, t^e_i \, < \, t^s_{k+1}\}$. See details in Appendix A.2.1.

Image Sharing & Image Reaction. The *image sharing & image reaction* functions are integrated to add a multi-modal dimension to the long-term dialogues.[2] The *image sharing* function is called when the agent decides to send an image. This

---

[2] Image captions are also saved to long-term memory.

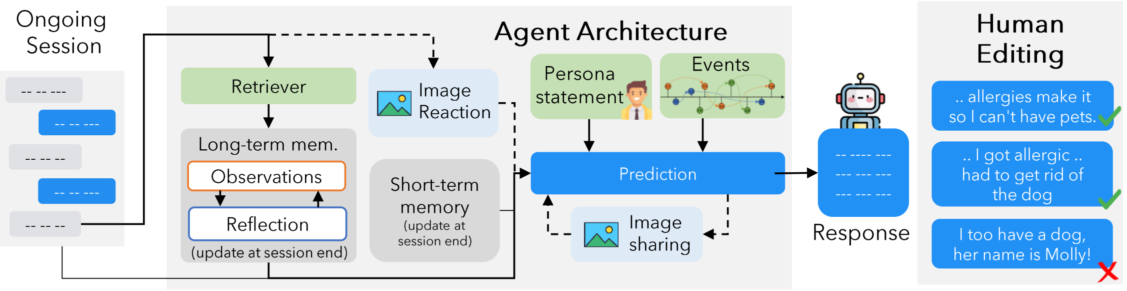


Figure 3: **Overview of the generative pipeline for LoCoMo.** Each LLM agent is assigned a distinct persona and a timeline of causally connected events in their file. The agent is equipped with a memory and reflection module to retrieve relevant history for dialog generation and is also enabled for image-sharing and image-reaction behaviors (left). The generated conversations are edited by human annotators to maintain long-range consistency (right).

process includes: (1) Generate a caption c for the intended image using $\mathcal{M}$; (2) Convert the caption c into relevant keywords w using $\mathcal{M}$; (3) Use the keywords k to find an image through web search $WEB(k)$[3]; (4) Share the chosen $image$. Conversely, the *image reaction* function is triggered upon receiving an image from another agent and entails: (1) Generate caption c for the received image[4]; (2) Generate a reaction for the received image in response using $\mathcal{M}$ (See Appendix A.2.1).

3.4 Human Verification & Editing

In the concluding phase, human annotators are tasked with (1) editing the dialogue to eliminate long-term inconsistencies, (2) removing or substituting irrelevant images, and (3) verifying and editing for alignment between event graphs and the content of the conversations. Overall, we observed that annotators edited nearly 15% of the dialog turns and removed or substituted approx. 19% images present in the LLM-generated dataset. See examples of some edits in Appendix A.3.

4 LoCoMo Evaluation Benchmark

Based on the dialogues generated in section 3, we introduce an evaluation benchmark (see Figure 2) composed of three tasks to assess the accuracy of *long-term memory*. See statistics of the dataset and evaluation benchmark in Table 5 in the Appendix.

4.1 Question Answering Task

A conversational agent is expected to possess a *memory* to remember previous dialogues, reflecting it to create more engaging responses in future conversations. For a comprehensive assessment of this *memory*, we introduce a question-answering task divided into five distinct reasoning categories: (1) **Single-hop** questions require answers based on a single session; (2) **Multi-hop** questions require synthesizing information from multiple different sessions; (3) **Temporal reasoning** questions can be answered through temporal reasoning and capturing time-related data cues within the conversation; (4) **Open-domain knowledge** questions can be answered by integrating a speaker's provided information with external knowledge such as commonsense or world facts; (5) **Adversarial** questions are designed to trick the agent into providing wrong answers, with the expectation that the agent will correctly identify them as unanswerable.

For each category, we calculate the F1 score for exact matches, following the normalization of both the predicted and the actual ground truth answers. However, evaluating long-form answers with automated metrics often presents challenges (Xu et al., 2023). LLMs tend to produce paraphrased responses in varied formats, complicating exact match evaluation. To simplify evaluation in our task, we ensure that answers in our QA annotations are directly taken from the conversations as much as possible. We instruct the LLMs to replicate the exact wording in the conversation when feasible and employ the F1 partial match metric for evaluating the predictions. Each QA sample is also annotated with the turn IDs in the conversation logs that contain the answer. We report the accuracy of retrieving the correct context for RAG models.

4.2 Event Summarization Task

The conversation is generated based on a temporal event graph $\mathcal{G}$ which is constructed by conditioning an LLM on a persona statement p, reflecting the chronological sequence of events in an individual's life. A conversational agent is expected to not only comprehend the causal connections and the

---

[3] https://pypi.org/project/icrawler/

[4] We use BLIP-2 (Li et al., 2023b) as the captioning model.

sequence of events in $\mathcal{G}$ but also to recount these events as required. To evaluate the agent's grasp of event dynamics, we introduce the event summarization task which challenges the agent to summarize the events within a designated timeframe and compares the agent's summary with events in $\mathcal{G}$. The events in LoCoMo are densely annotated lists of life events that are hard to summarize due to temporal and causal coreferences present in the dialogues, in contrast to existing summarization benchmarks of research papers (Li et al., 2023a), movie scripts (Chen et al., 2022), books (Kryściński et al., 2022), emails (Zhang et al., 2021b) etc.

Traditional metrics like BLEU (Papineni et al., 2002) and ROGUE (Lin, 2004) focus on lexical similarity between the reference and generated summaries, not meeting our needs as we emphasize factual accuracy in summarization. In this context, we employ FactScore (Min et al., 2023), a method that evaluates the factuality of generated text by decomposing both the reference and hypothesis into atomic facts. We adapt the metric to measure (1) *precision* of the summarized content by counting the number of atomic facts within the content that correspond with those in $\mathcal{G}$; (2) *recall* of the summarized content by determining how comprehensively the atomic facts of $\mathcal{G}$ are represented within the content. We present the F1 score, derived from the calculated precision and recall.

4.3 Multi-Modal Dialogue Generation Task

The conversations in our dataset are anchored to specific personas p and corresponding events $\mathcal{G}$ tailored to p. The topics in conversations evolve from events that were introduced in earlier dialogues, spanning weeks or months. This structure allows for an assessment of whether conversational agents can sustain a coherent persona and a continuous narrative over time. For example, if a speaker recently had an injury, the next conversations would likely focus on them recuperating, rather than engaging in adventurous activities. We assess such consistency by measuring how closely the predicted multi-modal dialogues align with the ground truth multi-modal dialogues in our dataset, quantifying this alignment through MMRelevance (Feng et al., 2022), in addition to other NLG metrics.

5 Experimental Setup

For the question-answering and event summarization tasks, we replace images in LoCoMo with their captions (Li et al., 2023b), and use state-of-art LLMs to reason over text-only dialogues interleaved with image captions. We use images directly for the multimodal dialog generation task only. See additional details in Appendix C.

Question Answering. We evaluate three types of models: (1) **Base** LLMs operating with constrained context lengths where earlier dialogues are omitted i.e., `Mistral-7B-Instruct-v0.2` (Jiang et al., 2023), `Llama-2-70b-chat` (Touvron et al., 2023), and `Llama-3-70B-Instruct`[5]; (2) **Long-context** LLMs with an extended context window i.e., `gpt-3.5-turbo` [6], `gpt-4-turbo` [7], `gemini-1.0-pro` (Team et al., 2023) and `claude-3-sonnet`[8]; (3) **Retrieval-augmented Generation (RAG)** involves retrieving relevant context from a database of dialog history, observations (assertions about speakers; see §3.3, Figure 9), or session-level summaries (see §3.3, Figure 8). We employ DRAGON (Lin et al., 2023) as retriever and `gpt-3.5-turbo` as reader.

Event Summarization. We present experiments using **Base** and **Long-context** setups from the question-answering task, but refrain from including RAG since summarization requires a comprehensive understanding of the entire dialogue, rather than just retrieving a specific portion. We implement incremental summarization i.e., iteratively create a summary of a preceding sessions and then use that summary as a basis to summarize the subsequent sessions (Chang et al., 2023).

Multi-modal Dialogue Generation. We generate 50 conversations using our automated pipeline (without human filtering; §3) for training data and train three versions of MiniGPT-5 (Zheng et al., 2023): (1) **Base** trains on prior dialogue turns only; (2) **+ summary** trains on prior dialogue turns and a global summary of the ongoing conversation; (3) **+ observation** trains on prior dialogue turns and observations retrieved from conversation history. Each run is initialized with a MiniGPT-5 checkpoint finetuned on MMDialog (Feng et al., 2022).

---

[5] https://llama.meta.com/llama3/

[6] https://platform.openai.com/docs/models/gpt-3-5

[7] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

[8] https://www.anthropic.com/news/claude-3-family

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-7B-Instruct-v0.2`	8K	19.1	15.1	9.3	8.6	28.9	18.7
	`Llama-2-70b-chat`	4K	**20.8**	**18.2**	**15.9**	**18.8**	15.7	18.4
	`Llama-3-70B-Instruct`	4K	17.0	17.0	12.0	13.0	**80.0**	**30.1**
Long context	`gpt-3.5-turbo`	4K	23.8	18.0	15.6	20.4	**34.8**	23.9
		8K	38.5	25.1	22.7	**25.9**	28.7	31.2
		12K	45.7	32.4	**25.5**	23.4	21.5	34.0
		16K	**52.6**	**36.7**	24.3	24.0	14.8	**35.9**
	`gemini-1.0-pro`	1M	62.4	35.3	34.2	19.0	5.2	39.1
	`claude-3-sonnet`	200K	70.7	38.1	26.9	52.2	2.5	42.8
	`gpt-4-turbo`	128K	**72.3**	**51.5**	**51.4**	**38.5**	**15.7**	**51.6**

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.

Retrieval Unit	top-k	Answer Prediction (F1 score)						Recall Accuracy (R@k)					
		Single Hop	Multi Hop	Temporal	Open Domain	Adver- -sarial	**Overall**	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	53.3	31.2	35.4	**25.0**	21.5	38.8	68.0	35.4	70.4	33.1	43.9	56.7
	10	56.9	34.6	34.5	23.9	**17.5**	39.7	77.7	46.6	77.3	40.8	54.7	66.2
	25	59.9	38.7	**37.2**	25.0	12.8	**41.0**	87.1	62.5	83.5	52.6	66.3	76.7
	50	**60.1**	**40.6**	36.9	22.4	9.9	40.5	**91.1**	**73**	**89.6**	**61.7**	**72.5**	**82.7**
`Observation`	5	54.3	36.3	**40.7**	**26.5**	**32.5**	**43.3**	67.1	41.4	73.1	35.1	37.7	56.2
	10	54.6	39.2	40.5	24.4	28.5	42.8	70.5	50.9	76.4	37.6	45.1	61.3
	25	**54.7**	41.2	38.8	25.8	24.7	42.1	74.7	60.6	80.3	49.0	53.5	67.5
	5	54.1	**41.6**	37.6	24.4	20.4	40.6	**76.3**	**67.8**	**82.0**	**55.9**	**59.5**	**71.2**
`Summary`	2	32.8	22.4	32.9	19.0	**25.3**	29.0	63.0	33.0	57.7	31.5	65.9	65.9
	5	35.1	26.0	37.4	21.2	23.5	30.9	77.0	54.3	72.8	47.6	79.1	72.1
	10	**36.0**	**29.9**	**37.5**	**22.2**	24.0	**32.0**	**88.7**	**72.7**	**84.5**	**67.3**	**88.8**	**84.7**

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

6 Experimental Results

We evaluate and analyze the comprehensive performance of all baseline methods for question answering (§6.1), event graph summarization (§6.2), and multi-modal dialogue generation (§6.3).

6.1 Question Answering Task

Tables 2 and 3 present the performance results for the question answering task. We find that: **(1) LLMs with limited context length face challenges in understanding extremely long conversations** due to truncated context windows. Despite `gpt-4-turbo` emerging as the top-performing model with an overall score of 51.6, it notably lags behind the human benchmark of 87.9; **(2) long-context LLMs can comprehend longer narratives, yet they are prone to generating hallucinations**. `gpt-4-turbo` outperforms other approaches on overall performance, but its performance on adversarial questions drops to a mere 15.7%, as compared to 34.8% using `gpt-3.5-turbo` and 80.0% using `llama-3-chat-70B` with 4K context lengths. Similar trends are observed in `gemini-pro-1.5` and `claude-sonnet` models as well. The overall performance of `gpt-3.5-turbo` increases with context length, mainly due to large improvements in single-hop and multi-hop scenarios, yet the performance on adversarial questions drops dramatically. This indicates that LLMs can be easily misled into generating hallucinations when they are subjected to long contexts; (3) **long-context LLMs struggle to utilize the recalled context correctly**. The performance gap between single-hop and multi-hop question categories demonstrates that LLMs have become reasonably good at 'memorizing' a large context window, but find it challenging to perform complex reasoning over the recalled context; **(4) RAG is effective when conversations are stored as observations**. There is a noticeable 5% improvement with `gpt-3.5-turbo` when the input is top 5 relevant observations instead of pure conversation logs. This improvement falters with an increase in the number of retrieved observations, suggesting that it is important to reduce the signal-to-noise (SNR) ratio in retrieved contexts for models to utilize the context accurately. Conversely, using session summaries as context does not significantly improve the performance despite high recall accuracies[9], likely due to loss of information during the conversion of dialogs to summaries.

The interesting finding is that **time reasoning and open-domain knowledge questions are the most challenging scenarios**. (1) LLMs face chal-

---

[9] For summary-based RAG models, the recall accuracy is based on retrieving the summary of the relevant session(s).

lenges in understanding time concepts within dialogues, which is consistent with findings from other single-turn-based benchmarks focused on temporal reasoning capabilities for LLMs (Wang and Zhao, 2023). (2) LLMs struggle with open-domain knowledge and degrade in the RAG setting. This suggests that while certain open-domain knowledge may be embedded within the model's parameters, introducing improper context from inaccurate retrieval can lead to a decline in performance (Mallen et al., 2023).

6.2 Event Summarization Task

Table 4 presents results for the event summarization task. Powerful long-context models record the highest performance on this task. `gpt-4-turbo` leads to the highest scores in terms of ROUGE and FactScore metrics, followed by `gemini-1.0-pro` and `claude-3-sonnet`. The use of incremental summarization with `Llama-3-70B-Instruct` (4K context window) performs reasonably well compared to the long-context models, demonstrating only a 2.4% drop in ROUGE-L score. However, there is a nearly 10% drop in performance on the FactScore metric, suggesting that it fails to capture as much information as the long-context models. Nonetheless, there remains considerable scope for improving performance on this task. The event summarization task requires long-range dependency to understand the temporal and causal connections between the events discussed by the speaker in multiple sessions (see Figure 7). The wide gap between the best-performing model and the upper bound on this task suggests that **long-context models may not be proficient at utilizing their context appropriately**, which also aligns with similar findings in Li et al. (2023a) as well as the QA task in LoCoMo.

From a manual analysis of predicted summaries, we identify five broad categories of event summarization errors made by LLMs: (1) **missing information** in events because the model fails to make temporal and/or causal connections over a lengthy conversation; (2) **hallucinations** i.e., models pad extra details that are either not present in the conversation or are part of a different event in the same session; (3) errors from **misunderstanding of dialog cues** such as humor or sarcasm is a distinctive issue with comprehension of dialogs; (4) inaccurate **speaker attributions**; and (5) insignificant dialogs that are wrongly considered as **salient** events. See examples in Table 6 in the Appendix.

6.3 Multi-Modal Dialog Generation Task

Figure 4 illustrates the effectiveness of various MiniGPT-5 training variants in multi-modal dialogue generation. Incorporating context into training enhances performance, with the inclusion of observation as context yielding significantly improved results. For instance, in Figure 4A, the retrieved observations contain information about the speaker's experience in video game tournaments, which leads to the prediction of dialog and images that are more faithful to the speaker's persona. This observation is consistent with earlier findings from the QA task as well (see Table 3). Also, we observe that the MM-Relevance score drops with an increase in the length of dialog history (see Figure 4B). Retrieval-augmented generation alleviates the drop in MM-Relevance to some extent.

7 Conclusion

We develop a human-machine pipeline to collect LoCoMo, a dataset of 10 high-quality very long conversations, each encompassing 600 turns and 16K tokens on avg., over up to 32 sessions, and propose an evaluation framework consisting of three tasks that evaluate models' proficiency in long conversations. Our experiments show that LLMs struggle to comprehend long-term narratives within the dialog and fail to draw temporal and causal connections between events discussed by speakers.

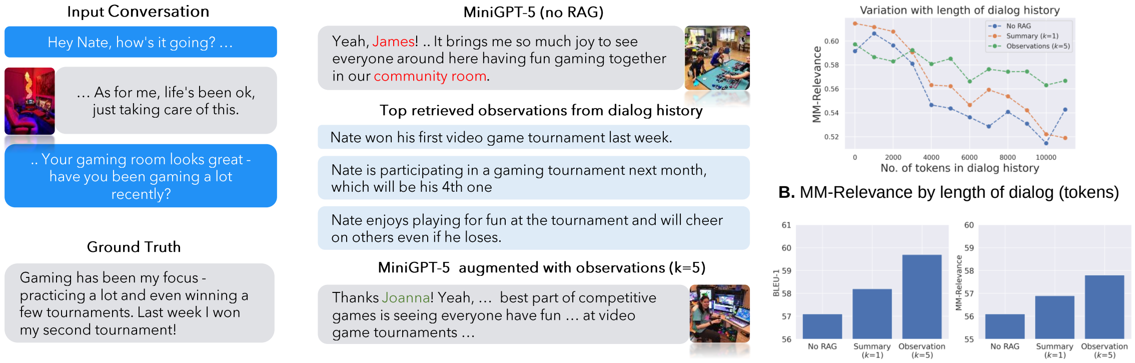
8 Limitations

Machine-generated data. Our dataset is sourced primarily from text generated by LLMs. We pursued this method, which has quickly emerged as a popular alternative to time-intensive manual data collection (Kim et al., 2023; Jang et al., 2023b), to avoid the logistical and legal complexities of collecting very long-term real-world conversations at scale. We ensure that the dataset mirrors real-world interactions as much as possible by having human annotators verify and edit the generated conversations. However, we acknowledge that this dataset may not fully reflect the nuances of real-world online conversations.

Limited exploration of multimodal behavior. Since the images in our dataset are sourced from the web, they do not demonstrate the visual long-term consistencies that are usually exhibited in personal photos (e.g., appearance, home environment, people and pets, etc.). Consequently, we find that the

Category	Model	Context Length	ROUGE			FactScore		
			ROUGE-1	ROUGE-2	ROUGE-L	Precision	Recall	F1
Base	`Mistral-7B-Instruct-v0.2`	8K	34.6	10.1	16.4	33.5	31.2	32.3
	`Llama-3-70B-Instruct`	4K	36.7	11.4	19.2	40.3	35.6	37.8
Long context	`gemini-1.0-pro`	1M	37.6	13.4	21.1	46.7	42.1	44.2
	`claude-3-sonnet`	200K	35.1	12.6	21.3	45.6	40.8	43.1
	`gpt-4-turbo`	128K	**41.2**	**13.8**	**21.6**	**51.9**	**46.5**	**48.9**

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.



Hey Nate, how's it going? …

… As for me, life's been ok, just taking care of this.

.. Your gaming room looks great - have you been gaming a lot recently?

Gaming has been my focus - practicing a lot and even winning a few tournaments. Last week I won my second tournament!

Yeah, James! .. It brings me so much joy to see everyone around here having fun gaming together in our community room.

Nate won his first video game tournament last week.

Nate is participating in a gaming tournament next month, which will be his 4th one

Nate enjoys playing for fun at the tournament and will cheer on others even if he loses.

Thanks Joanna! Yeah, … best part of competitive games is seeing everyone have fun … at video game tournaments …

B. MM-Relevance by length of dialog (tokens)

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

C. BLEU-1, MM-Relevance of various methods

Figure 4: **Multimodal dialog generation performance of MiniGPT-5**. (A) an example of multimodal dialog predicted using MiniGPT5 with and without observation as retrieved context, (B) Variation of MM-Relevance score with length of dialog history, and (C) comparison of RAG-based MiniGPT-5 methods.

images in our dataset can be replaced with their captions without much loss of information, except for cases where OCR is required. Nevertheless, our work is a first step toward research into the multimodal aspect of very long-term conversations.

Language. Our LLM-based pipeline for generating long-term conversations has been developed for the English language only. However, our pipeline can be made to work with any other language using an LLM that is proficient at that language and appropriate translations of our prompts.

Closed-source LLMs. We use state-of-the-art LLMs in our dialog generation pipeline to create a dialog dataset that is as realistic as possible. Unfortunately, this meant employing LLMs that are not open-sourced and are only available through a paid API, similar to many concurrent works that generate synthetic conversations (Zhong et al., 2023; Lu et al., 2023). We will make the code for our generative pipeline publicly available in the hope that it can be made to work effectively with open-source LLMs in the future.

Evaluation of long-form NLG. LLMs are prone to generating verbose answers even when prompted to answer in short phrases. This creates challenges in evaluating the correctness of answers provided by LLMs and has been widely documented in NLP literature (Chang et al., 2023; Xu et al., 2023; Krishna et al., 2023). Our evaluation framework suffers from the same challenges when used for experimenting with LLMs.

9 Broader Impacts

We adopt and improve a framework of generative agents introduced in Park et al. (2023) for the generation of long-term conversations. Consequently, the ethical concerns of generative agents outlined by Park et al. (2023) apply to our work as well, especially since the goal of our framework is to make the conversations as realistic as possible.

Specifically, conversational agents that can pose as human beings with a realistic life, as enabled by the temporal event graphs in our framework, pose the risk that users may form parasocial relationships with such agents that may affect their lives adversely. We recommend that any practical deployment of the generative framework in our work be always prefaced with a disclaimer about the source of the dialogs.

Second, the use of multimodal LLMs (Zheng et al., 2023) to generate images conditioned on dialog can lead to the propagation of misinformation and social biases, especially if the conversational agent can be coerced into parroting false information or dangerous opinions.

Third, it is tempting to use generative agents to substitute real humans for a process, especially when there are significant challenges in working with humans for a particular goal e.g., collecting real-world interactions between humans over a year or more. Care must be taken to ensure that such substitutes are not made in studies whose outcomes may be used to make real-world decisions with tangible impacts on humans. Our work is merely a study of model comprehension in very long-term conversations. We do not make any recommendations for real-world policies based on this study and advise potential users of our framework to avoid making such recommendations as well.

References

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpchat: Towards multimodal persona-grounded conversation. *arXiv preprint arXiv:2305.17388*.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jan Assmann and John Czaplicka. 1995. Collective memory and cultural identity. *New german critique*, (65):125–133.

Amanda Bertsch, Uri Alon, Graham Neubig, and Matthew R Gormley. 2023. Unlimiformer: Long-range transformers with unlimited length input. *arXiv preprint arXiv:2305.01625*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. Summscreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*.

Alan Cooper. 1999. *The inmates are running the asylum*. Springer.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023a. Peacok: Persona commonsense knowledge for consistent and engaging narratives. *arXiv preprint arXiv:2305.02364*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. Deam: Dialogue coherence evaluation using amr-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785.

Xingwei He, Zhenghao Lin, Yeyun Gong, Alex Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854*.

William Hirst and Gerald Echterhoff. 2012. Remembering in conversations: The social sharing and reshaping of memories. *Annual review of psychology*, 63:55–79.

William Hirst and David Manier. 2008. Towards a psychology of collective memory. *Memory*, 16(3):183–200.

William Hirst, Jeremy K Yamashiro, and Alin Coman. 2018. Collective memory from a psychological perspective. *Trends in cognitive sciences*, 22(5):438–451.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023a. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. *arXiv preprint arXiv:2310.13420*.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023b. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13584–13606, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1642–1661.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558.

Dong-Ho Lee, Jay Pujara, Mohit Sewak, Ryen White, and Sujay Jauhar. 2023a. Making large language models better data creators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15349–15360, Singapore. Association for Computational Linguistics.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023b. Prompted llms as chatbot modules for long open-domain conversation. *arXiv preprint arXiv:2305.04533*.

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, Jonghwan Hyeon, and Ho-Jin Choi. 2023c. Dialogcc: An automated pipeline for creating high-quality multimodal dialogue datasets. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023a. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Xinnian Liang, Bing Wang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023. Unleashing infinite-length input capacity for large-scale language models with self-controlled memory system. *arXiv preprint arXiv:2304.13343*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your dragon: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400, Singapore. Association for Computational Linguistics.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023b. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023.

Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

John Pruitt and Jonathan Grudin. 2003. Personas: practice and theory. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Michael Shum, Stephan Zheng, Wojciech Kryściński, Caiming Xiong, and Richard Socher. 2019. Sketch-fill-ar: A persona-grounded chit-chat generation framework. *arXiv preprint arXiv:1910.13008*.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yuqing Wang and Yun Zhao. 2023. Tram: Benchmarking temporal reasoning for large language models. *arXiv preprint arXiv:2310.00835*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. *arXiv preprint arXiv:2305.18201*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453*.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021a. Dynaeval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022. Finedeval: Fine-grained automatic dialogue-level evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3336–3355.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Mind the gap between conversations for improved long-term dialogue generation. *arXiv preprint arXiv:2310.15415*.

Shiyue Zhang, Asli Celikyilmaz, Jianfeng Gao, and Mohit Bansal. 2021b. Emailsum: Abstractive email thread summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6895–6909.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2021. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, and Yanlin Wang. 2023. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Overview

The appendix is organized as follows:
Section A: Details of generative pipeline for the LoCoMo dataset.
Section B: Statistics of LoCoMo dataset, license for data release and annotator details.
Section C: Experimental setup and implementation details.
Section D: Additional results from evaluation on the LoCoMo benchmark.

A Generative Pipeline for LoCoMo

A.1 Persona

We assign unique persona statement p to each agent $\mathcal{L}_i$. For this, we select a range of initial persona statements p_c from the MSC dataset (Xu et al., 2022), each encompassing 4 to 5 sentences. We employ `gpt-3.5-turbo` as $\mathcal{M}$ to expand these into full persona statement p, conditioning $\mathcal{M}$ on the chosen statements p_c. The prompt used for converting a short list of speaker attributes from the MSC dataset (Xu et al., 2022) into a complete persona summary is presented in Fig. 5. We also use a single example of speaker attribute → persona summary as an in-context demonstration along with the prompt. A small selection of personas showcasing the diversity of speakers in the LoCoMo dataset is demonstrated in Fig. 5.

A.2 Temporal Event Graph

As outlined in Sec. 3.2, we use an iterative process for generating event graphs consisting of causally connected events based on a given persona summary. The base prompt for describing the constitution of the event graph, the nature of events and causal connections between events is shown in Fig. 6. First, the base prompt is used along with the prompt for event graph initialization to generate three independent events relevant to a given personality. Then, the base prompt is combined with the prompt for the iterative generation of events to continue generating events that are caused by one or more of the events that are already present in the graph. See an example of a persona and the corresponding temporal event graph in Fig. 7. In the example, Jack aspires to be a hotel manager. Consequently, he enrolls in a hotel management course in July, and after three months, he expresses his excitement about the course on social media. In a similar vein, his passion for gaming results in an invitation from a well-known gaming company.

A.2.1 Virtual Agent Architecture

As outlined in Section 3.3, the virtual agents in our generative pipelines are composed of two mechanisms, *Reflect & respond* (Park et al., 2023) and *Image sharing & response*.

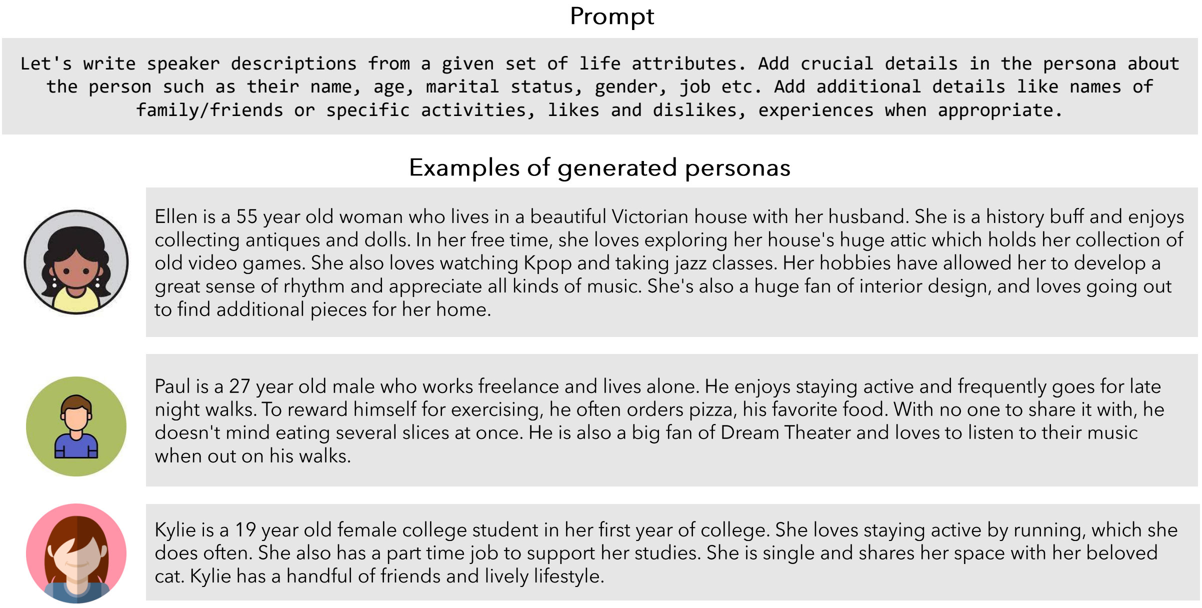
Reflect & respond. This mechanism operates over a combination of short-term and long-term memory. The short-term memory is a summary of a session that is conditioned on the summary from a previous session. See the prompt given to LLMs in our pipeline for generating summaries, and an example of a generated summary, in Fig. 8. The long-term memory is a database of *observations* about each speaker, that are essentially assertive statements about the speaker's persona and life. See the prompt given to LLMs in our pipeline for generating observations, and an example of observations extracted from a conversation, in Fig. 9. In practice, the conversation is annotated with turn IDs for each turn, and the model is also instructed to indicate the turn IDs that directly contribute to each observation. This allows us to keep track of the evidence when using observations as the context for RAG-based models used in our experiments (see Section 5).

Image sharing & response. See prompts for implementing image-sharing and image-response behaviors in Figure 10.

A.3 Human Filtering

Human annotators are instructed to edit the LLM-generated conversations in the following scenarios:

- Remove an image if it is not relevant to the current dialog or the conversation.
- Add context about an image to the current speaker's dialog if it is not discussed by them but the subsequent speaker has reacted to the image.
- Replace an image if it does not match the caption that was used to query for images.
- Edit the dialog when the information present in the dialog is inconsistent with something said (or shared through an image) in earlier or later turns.



Prompt

Let's write speaker descriptions from a given set of life attributes. Add crucial details in the persona about the person such as their name, age, marital status, gender, job etc. Add additional details like names of family/friends or specific activities, likes and dislikes, experiences when appropriate.

Examples of generated personas

Ellen is a 55 year old woman who lives in a beautiful Victorian house with her husband. She is a history buff and enjoys collecting antiques and dolls. In her free time, she loves exploring her house's huge attic which holds her collection of old video games. She also loves watching Kpop and taking jazz classes. Her hobbies have allowed her to develop a great sense of rhythm and appreciate all kinds of music. She's also a huge fan of interior design, and loves going out to find additional pieces for her home.

Paul is a 27 year old male who works freelance and lives alone. He enjoys staying active and frequently goes for late night walks. To reward himself for exercising, he often orders pizza, his favorite food. With no one to share it with, he doesn't mind eating several slices at once. He is also a big fan of Dream Theater and loves to listen to their music when out on his walks.

Kylie is a 19 year old female college student in her first year of college. She loves staying active by running, which she does often. She also has a part time job to support her studies. She is single and shares her space with her beloved cat. Kylie has a handful of friends and lively lifestyle.

Figure 5: **Prompt for persona statement (p) generation and examples of personas in LoCoMo.** The prompt used to generate expanded persona statements (p) from initial personas (p_c) for the virtual agents in our conversation generation pipeline (top) and select examples of persona statements present in the LoCoMo dataset.

- Edit the dialog to ensure that the details in the conversation are consistent with those given in the event for the session.
- Remove any events from the event graph if they do not appear in the conversation.

See an example of some edits in Fig. 11.

B Dataset

B.1 Dataset Statistics

See a breakdown of the statistics of the conversations in the LoCoMo dataset in the top panel of Table 5. Also, see a breakdown of the statistics of the annotations in the evaluation benchmark in the bottom panel of Table 5.

B.2 Dataset License

The LoCoMo dataset will be released under the CC BY-NC 4.0 DEED license.[10]

B.3 Annotator Details

The annotators who worked on the LoCoMo dataset were in-house annotators and we were unable to obtain their demographics due to the confidential nature of such information.

Conversation Statistics	# Counts
Total. # conversations h.	10
Avg. # sessions k. in conversation h	27.2
Avg. # turns j. in session k	21.6
Avg. # tokens. conversation h	16,618.1
Avg. # tokens. dialogue h_{k_j} of turn j in session k	29.8
Avg. # tokens. observation o_{k_j} of turn j in session k	19.2
Avg. # tokens. summary w_k of session k	132.4
QA Benchmark Statistics	
# questions. single-hop retrieval	841 (42.3%)
# questions. multi-hop retrieval	282 (14.2%)
# questions. temporal reasoning	321 (16.1%)
# questions. open domain knowledge	96 (4.8%)
# questions. adversarial	446 (22.4%)
Total. # questions.	**1,986**
Event Summarization Statistics	
Avg. # ground truth events. in conversation h	35.8
Avg. # tokens. event summary	1042.7
Multi-modal Dialogue Generation Statistics	
Avg. # images. in conversation h	91.2

Table 5: **Dataset Statistics** of conversation and corresponding benchmark

C Experimental Setup

C.1 Baselines

The conversations in the LoCoMo dataset are composed of natural language dialogs and images that require higher-order reasoning and multimodal coreference resolution, respectively. From initial studies, we observed that multimodal coreference resolution can be performed effectively by replacing images in LoCoMo with their captions gen-

---

[10] https://creativecommons.org/licenses/by-nc/4.0/

Base prompt for event graph generation

```
Let's write a graph representing events that occur in a person's life based on a short summary of their
personality. Nodes represent events and edges represent the influence of past sub-events on a current event.
- The graph is represented in the form of a json list.
- Each entry is a dictionary containing the following keys: "event", “date", "caused_by", "id".
- The "event" field contains a short description of the event.
- The “date" field contains a date.
- The "id" field contains a unique identifier for the event.
- The "caused_by" field represents edges and is a list of "id" of existing events that have caused this event.
Events in the "caused_by" field should occur on dates before the event they have caused. Generate as many
causal connections as possible.
- An example of a causal effect is when the event "started a vegetable garden" causes "harvested tomatoes".
- Events can be positive or negative life events.
```

Additional prompt for event graph initialization

```
For the following input personality, generate three independent events E1, E2 and E3 aligned with their
personality. Events can be positive or negative life events and should reflect evolution in the person's
relationships, state of mind, personality etc.
```

Additional prompt for iterative generation of causal events

```
For the following input personality, generate new events that are caused by one or more EXISTING events. Events
can be positive or negative life events and should reflect evolution in the person's relationships, state of
mind, personality etc. Do not repeat existing sub-events. Start and end your answer with a square bracket.
```

Figure 6: **Prompts for temporal event graph generation.** The prompt used to generate complete personas for the LLMs in our conversation generation pipeline (top) and examples of personas present in the LoCoMo dataset.

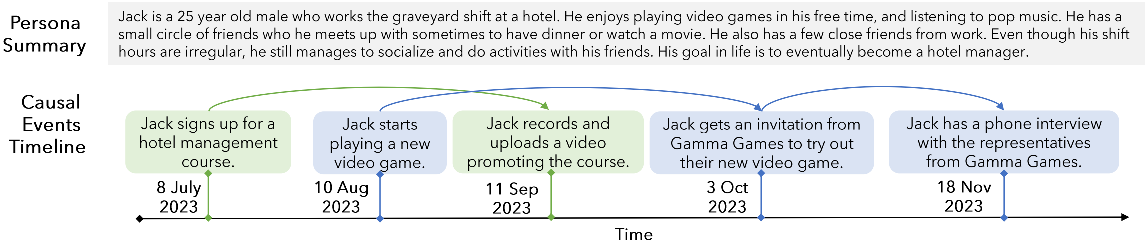
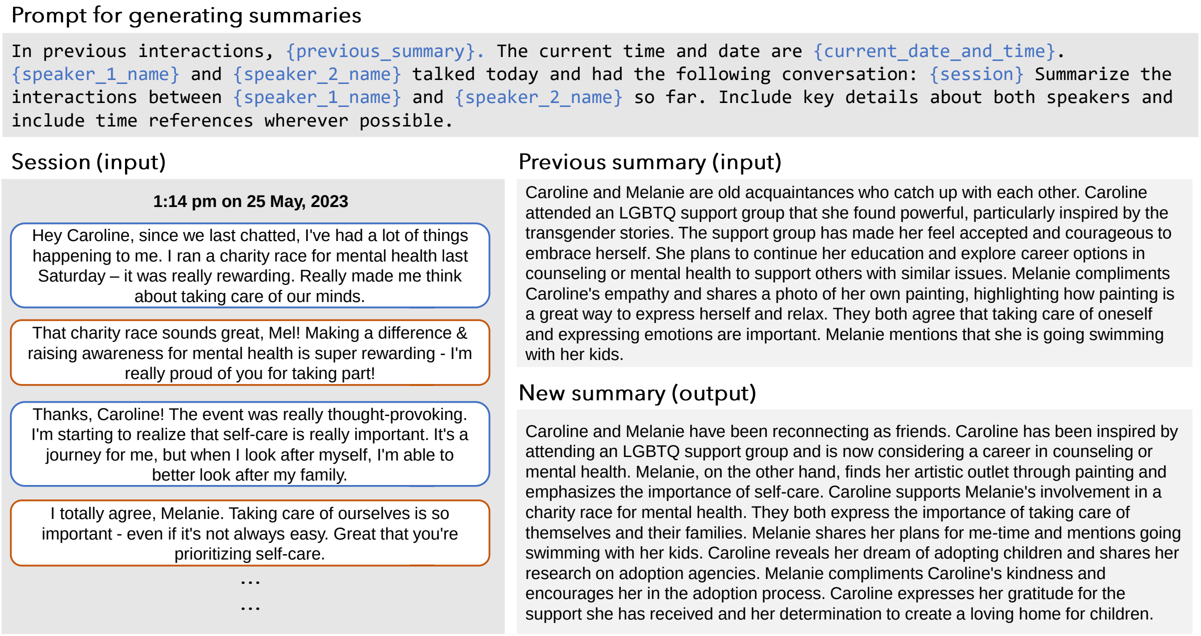

Jack is a 25 year old male who works the graveyard shift at a hotel. He enjoys playing video games in his free time, and listening to pop music. He has a small circle of friends who he meets up with sometimes to have dinner or watch a movie. He also has a few close friends from work. Even though his shift hours are irregular, he still manages to socialize and do activities with his friends. His goal in life is to eventually become a hotel manager.
Persona Summary
Causal Events Timeline
Jack signs up for a hotel management course.
8 July 2023
Jack starts playing a new video game.
10 Aug 2023
Jack records and uploads a video promoting the course.
11 Sep 2023
Jack gets an invitation from Gamma Games to try out their new video game.
3 Oct 2023
Jack has a phone interview with the representatives from Gamma Games.
18 Nov 2023
Time

Figure 7: **Temporal Event Graph $\mathcal{G}$ Creation.** Each event is generated in accordance with the specified persona p and causal connections l between events are depicted to illustrate the casual relationships among them.

erated using BLIP-2 (Li et al., 2023b), and using state-of-art LLMs to reason over natural language text interleaved with image captions. Hence, our experiments for the question answering and event summarization tasks are conducted using LLMs. We use the images directly only for experiments on the multimodal dialog generation task.

Question Answering. We carry out experiments using three distinct methodologies: (1) **Base** involves utilizing LLMs to directly conduct the task within a constrained context. The task description comes after the dialogue history. To accommodate the restricted context window size, earlier dialogues are omitted; (2) **Long-context** employs LLMs with an extended context window to expose the models to as much dialogue context as possible; (3) **Retrieval-augmented Generation (RAG)** involves retrieving relevant context from a database of dialog history, observations, or session-level summaries. *Observations* are assertions about each speaker extracted from the dialog history as described in §3.3, see an example in Figure 9. Session-level *summaries* are concise summaries of the conversation that takes place in each session, see an example in Figure 8. For the retrieval model, we employ DRAGON (Lin et al., 2023). To assess the effectiveness in practical scenarios for *Long-context* and *RAG*, we draw comparisons using variants of `gpt-3.5-turbo`. We do not report the performance of long-context fine-tuned open-source models (Chen et al., 2023) or those utilizing sliding window (Bertsch et al., 2023; Dao et al., 2022) due to the variability inherent across different open-source models and the potential reduction in their capability on shorter context.

Event Summarization. We present experiments conducted in two distinct configurations. We use both the **Base** and **Long-context** setups from the



Prompt for generating summaries

```
In previous interactions, {previous_summary}. The current time and date are {current_date_and_time}.
{speaker_1_name} and {speaker_2_name} talked today and had the following conversation: {session} Summarize the
interactions between {speaker_1_name} and {speaker_2_name} so far. Include key details about both speakers and
include time references wherever possible.
```

Session (input)

1:14 pm on 25 May, 2023

Hey Caroline, since we last chatted, I've had a lot of things happening to me. I ran a charity race for mental health last Saturday – it was really rewarding. Really made me think about taking care of our minds.

That charity race sounds great, Mel! Making a difference & raising awareness for mental health is super rewarding - I'm really proud of you for taking part!

Thanks, Caroline! The event was really thought-provoking. I'm starting to realize that self-care is really important. It's a journey for me, but when I look after myself, I'm able to better look after my family.

I totally agree, Melanie. Taking care of ourselves is so important - even if it's not always easy. Great that you're prioritizing self-care.

…

…

Previous summary (input)

Caroline and Melanie are old acquaintances who catch up with each other. Caroline attended an LGBTQ support group that she found powerful, particularly inspired by the transgender stories. The support group has made her feel accepted and courageous to embrace herself. She plans to continue her education and explore career options in counseling or mental health to support others with similar issues. Melanie compliments Caroline's empathy and shares a photo of her own painting, highlighting how painting is a great way to express herself and relax. They both agree that taking care of oneself and expressing emotions are important. Melanie mentions that she is going swimming with her kids.

New summary (output)

Caroline and Melanie have been reconnecting as friends. Caroline has been inspired by attending an LGBTQ support group and is now considering a career in counseling or mental health. Melanie, on the other hand, finds her artistic outlet through painting and emphasizes the importance of self-care. Caroline supports Melanie's involvement in a charity race for mental health. They both express the importance of taking care of themselves and their families. Melanie shares her plans for me-time and mentions going swimming with her kids. Caroline reveals her dream of adopting children and shares her research on adoption agencies. Melanie compliments Caroline's kindness and encourages her in the adoption process. Caroline expresses her gratitude for the support she has received and her determination to create a loving home for children.

Figure 8: **Prompt for generating conversation summaries.** The prompt used to iteratively generate a summary for the current session by conditioning on summary from preceding sessions and the raw conversation logs of the current session (top); and an example of inputs for the prompt and corresponding output summary of a session from the LoCoMo dataset.

question answering task, but we refrained from including RAG since summarization requires a comprehensive understanding of the entire dialogue, rather than just retrieving a specific portion. A notable distinction in our approach, compared to the question-answering task, lies in our handling of the context. Specifically, we employ an iterative process of creating a summary of a preceding session and then use that summary as a basis to generate the summary for the subsequent session (Chang et al., 2023). Further, we use a single in-context demonstration of input and output to guide the model toward selecting only significant life events for the summary.

Multi-modal Dialogue Generation. For evaluating multi-modal dialogue generation, we train MiniGPT-5 (Zheng et al., 2023) on 50 conversations generated using our automated pipeline (without human filtering) as detailed in §3. Three distinct versions of the model were developed, each with varying training data: (1) **Base** trains on preceding dialogue turns; (2) **+ summary** trains on both prior dialogue turns and a global summary of the ongoing conversation; (3) **+ observation** trains on both preceding dialogue turns and relevant observations retrieved from the conversation history. For each of these models, we started with a MiniGPT-5 checkpoint pretrained on the MMDialog dataset (Feng et al., 2022).

C.2 Implementation Details

We use OpenAI API, Gemini API, Claude API and Huggingface (Wolf et al., 2020), as of May 2024, with specific settings of $temperature$ set to 0 and top_p set to 1 for evaluation of the LoCoMo benchmark. All experiments, including those for RAG-based models, MiniGPT-5 training, and inference, are conducted on an Nvidia A6000 server with FP32. We report results from a single inference run for each model in our experiments. For MiniGPT-5, we used the hyperparameters recommended in the original codebase and trained our models for 10 epochs, which took approximately 30 hours on a single A6000 GPU.

We use the default implementations of BLEU[11], ROUGE[12], BertScore[13], FactScore[14] metrics in their respective Python packages in our evaluation protocol.

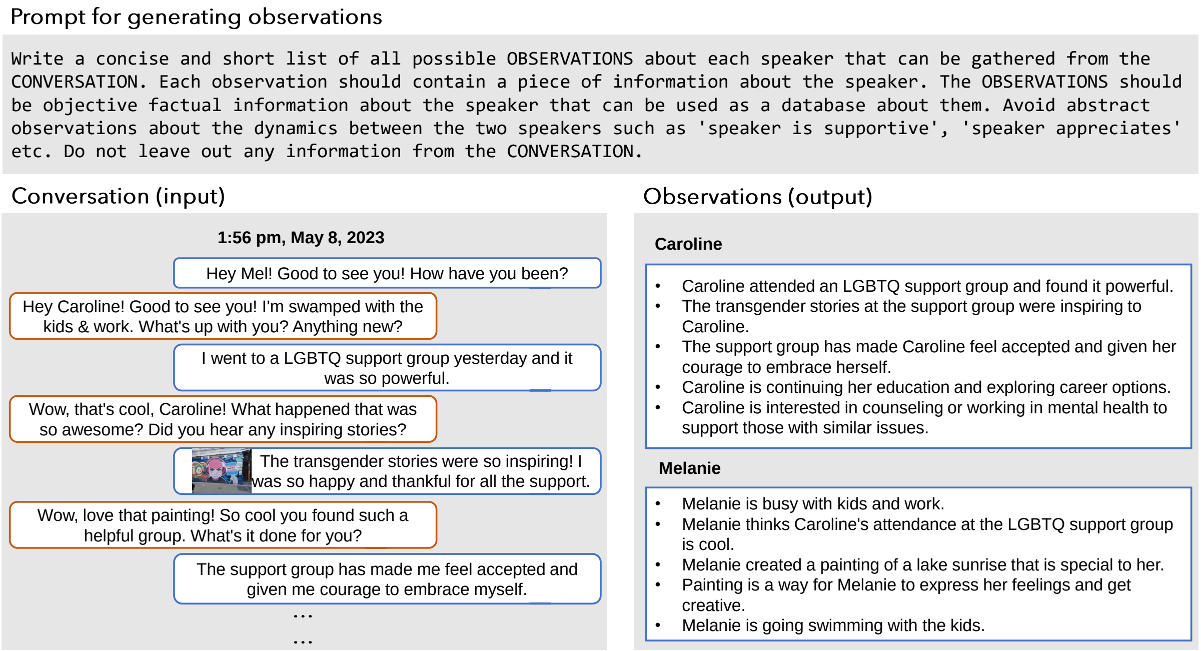
---

[11] https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[12] https://pypi.org/project/rouge/

[13] https://pypi.org/project/bert-score/

[14] https://github.com/shmsw25/FActScore



Prompt for generating observations

```
Write a concise and short list of all possible OBSERVATIONS about each speaker that can be gathered from the
CONVERSATION. Each observation should contain a piece of information about the speaker. The OBSERVATIONS should
be objective factual information about the speaker that can be used as a database about them. Avoid abstract
observations about the dynamics between the two speakers such as 'speaker is supportive', 'speaker appreciates'
etc. Do not leave out any information from the CONVERSATION.
```

Conversation (input)

1:56 pm, May 8, 2023

Hey Mel! Good to see you! How have you been?

Hey Caroline! Good to see you! I'm swamped with the kids & work. What's up with you? Anything new?

I went to a LGBTQ support group yesterday and it was so powerful.

Wow, that's cool, Caroline! What happened that was so awesome? Did you hear any inspiring stories?

The transgender stories were so inspiring! I was so happy and thankful for all the support.

Wow, love that painting! So cool you found such a helpful group. What's it done for you?

The support group has made me feel accepted and given me courage to embrace myself.

…

…

Observations (output)

Caroline

- Caroline attended an LGBTQ support group and found it powerful.
- The transgender stories at the support group were inspiring to Caroline.
- The support group has made Caroline feel accepted and given her courage to embrace herself.
- Caroline is continuing her education and exploring career options.
- Caroline is interested in counseling or working in mental health to support those with similar issues.

Melanie

- Melanie is busy with kids and work.
- Melanie thinks Caroline's attendance at the LGBTQ support group is cool.
- Melanie created a painting of a lake sunrise that is special to her.
- Painting is a way for Melanie to express her feelings and get creative.
- Melanie is going swimming with the kids.

Figure 9: **Prompts for generating observations from conversations.** The prompt used to generate observations from a conversation (top); and an example of inputs for the prompt and corresponding output observations for a session from the LoCoMo dataset.

D Results

D.1 Event Summarization Task

See an example of the five broad categories of event summarization errors made by LLMs, outlined in Section 6.2, in Table 6.

D.2 Multimodal Dialog Generation Task

Results from evaluation of various version of MiniGPT-5 model on the multimodal dialog generation task in the LoCoMo benchmark is shown in Table 7.

Image sharing: Prompt for generated image caption → image query

```
Let's write short image search queries from textual descriptions of photos shared by a user.
Queries should not include names of people, years and other irrelevant details. For example:

Input: That sounds relaxing, Jeremy! As for video game suggestions, have you ever tried "The Legend
of Zelda: Breath of the Wild"? It's an open-world adventure game that I absolutely love. [shares a
photo of Link standing in front of a breathtaking landscape] Have a look at this stunning view!
Output: the legend of zelda: breath of wild link landscape

Input: {generated_image_caption}
Output:
```

Image response: Prompt for shared image caption → response grounded in image

```
{speaker_1_persona}
{speaker_2_persona}

{speaker_1_name} says, {current_turn}, and shares a photo of {shared_image_caption_blip2}. Write
the most natural question or comment {speaker_2_name} can include in their response.
```

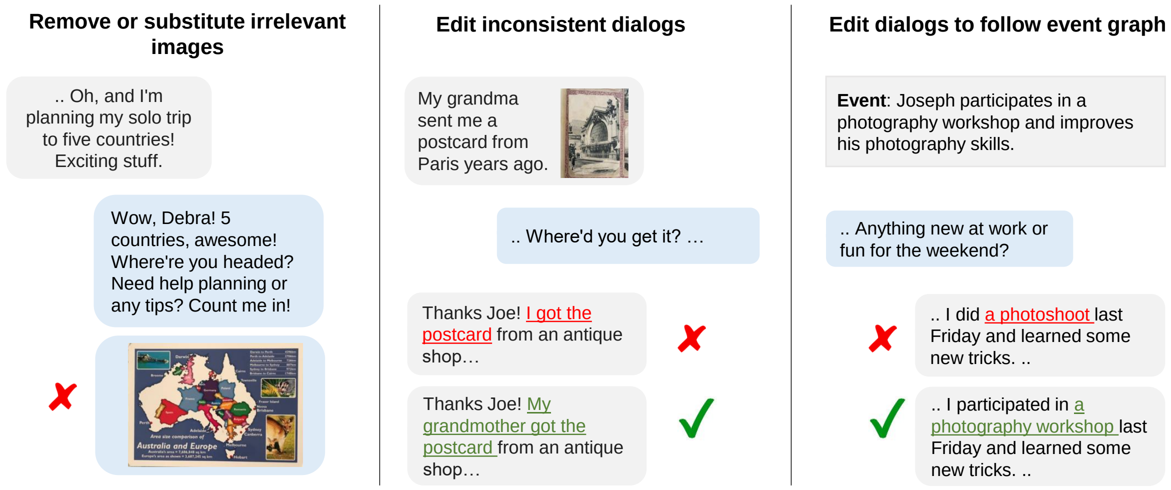
Figure 10: **Prompts for image-sharing and image-response behavior.** The prompt used to convert a caption generated by the virtual agent into an image query for the web-based image crawler in our pipeline (top), and the prompt used to generate a response grounded in the image shared by a virtual agent during a conversation as well as the personas of the respective speakers (bottom).

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 6: **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`.

Category	**top-k**	**BLEU-1/2**	**Rouge-L**	**MM-R**
`Base`	-	56.4 / 31.8	11.6	54.2
`+ summary`	1	57.2 / 31.6	11.9	54.7
`+ summary`	2	56.6 / 30.9	11.7	54.1
`+ summary`	5	56.2 / 30.5	11.5	54.0
`+ observation`	5	**58.7 / 32.2**	**12.6**	**55.8**
`+ observation`	10	58.1 / 32.1	12.0	55.1
`+ observation`	25	57.8 / 31.6	11.8	54.9

Table 7: **Multi-modal dialogue generation performance** comparison between different training variants of MiniGPT-5. The optimal performance is shown in **bold**.


Remove or substitute irrelevant images
Edit inconsistent dialogs
Edit dialogs to follow event graph
✗
✗
✓
✗
✓

.. 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!

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…

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. ..

Figure 11: **Example of edits made by annotators.** Human annotators are instructed to make edits in the LLM-generated conversations to remove irrelevant The prompt used to generate complete personas for the LLMs in our conversation generation pipeline (top) and examples of personas present in the LoCoMo dataset.