SeCom: On Memory Construction and Retrieval for Personalized Conversational Agents

Zhuoshi Pan[1,*] **Qianhui Wu**[2,†] **Huiqiang Jiang**[2], **Xufang Luo**[2], **Hao Cheng**[2], **Dongsheng Li**[2], **Yuqing Yang**[2], **Chin-Yew Lin**[2], **H. Vicky Zhao**[1], **Lili Qiu**[2], **Jianfeng Gao**[2]
[1] Tsinghua University, [2] Microsoft Corporation
{qianhuiwu, hjiang, xufang.luo}@microsoft.com

Abstract

To deliver coherent and personalized experiences in long-term conversations, existing approaches typically perform retrieval augmented response generation by constructing memory banks from conversation history at either the turn-level, session-level, or through summarization techniques. In this paper, we explore the impact of different memory granularities and present two key findings: (1) Turn-level, session-level, and summarization-based methods all exhibit limitations in terms of the accuracy of the retrieval and the semantics of the retrieved content, ultimately leading to sub-optimal responses. (2) The redundancy in natural language introduces noise, hindering precise retrieval. We demonstrate that *LLMLingua-2*, originally designed for prompt compression to accelerate LLM inference, can serve as an effective denoising method to enhance memory retrieval accuracy.

Building on these insights, we propose **SECOM**, a method that constructs the memory bank at segment level by introducing a **SE**gmentation model that partitions long-term conversations into topically coherent segments, while applying **COM**pression based denoising on memory units to enhance memory retrieval. Experimental results show that SECOM exhibits superior performance over baselines on long-term conversation benchmarks *LOCOMO* and *Long-MT-Bench+*.

1 Introduction

Large language models (LLMs) have developed rapidly and have been widely used in conversational agents. In contrast to traditional dialogue systems, which typically focus on short conversations within specific domains [1], LLM-powered conversational agents engage in significantly more interaction turns across a broader range of topics in open-domain conversations [2, 3]. Such long-term, open-domain conversations over multiple sessions present significant challenges, as they require the system to retain past events and user preferences to deliver coherent and personalized responses [4].

Some methods maintain context by concatenating all historical utterances or their summarized versions [5, 6]. However, these strategies can lead to excessively long and irrelevant contexts that distract the LLM, hindering its comprehension of the conversation, as noted by Maharana et al. [7]. Some other works focus on retrieving query-related conversation history to enhance response generation [8–10, 7]. These approaches typically construct memory bank from the conversation history at either the *turn-level* [8] or *session-level* [6]. Xu et al. [11], Chen et al. [4] and Zhong et al. [12] further leverage *summarization* techniques to build memory units, which are then retrieved as context for response generation.

[1]Work during internship at Microsoft.
[2]Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Building on these works, a key question arises: Which level of memory granularity—turn-level, session-level, or their summarized forms—yields the highest effectiveness? Moreover, is there a novel memory structure that could outperform these three formats?

In this paper, we first systematically investigate the impact of different memory granularities on conversational agents within the paradigm of retrieval augmented response generation [13, 14]. Our findings indicate that turn-level, session-level, and summarization-based methods all exhibit limitations in terms of the retrieval accuracy as well as the semantics of the retrieved content, which ultimately lead to sub-optimal responses, as depicted in Figure 1, Figure 2, and Table 1.

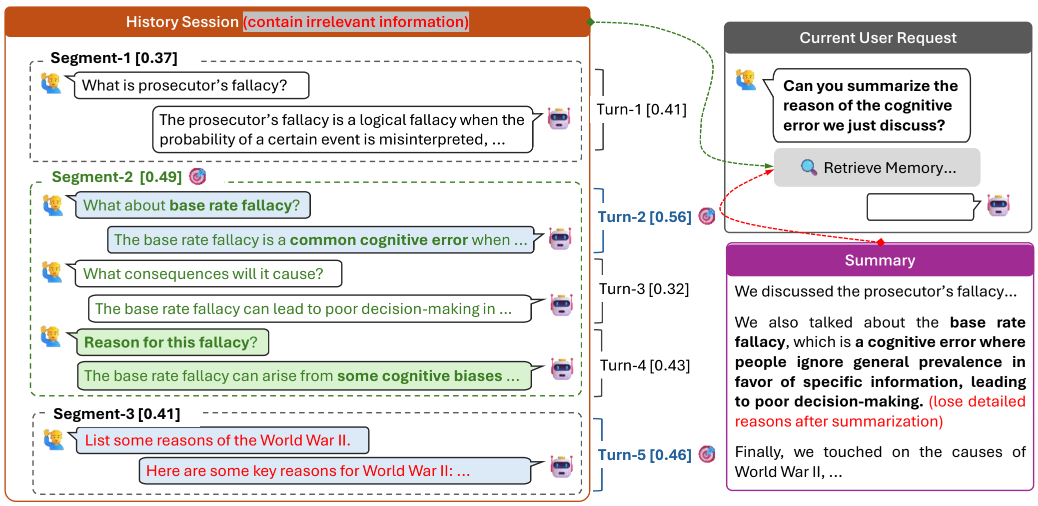
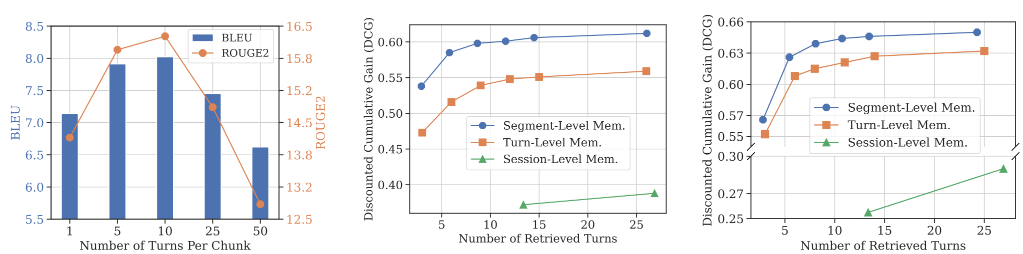


Figure 1: Illustration of retrieval augmented response generation with different memory granularities. *Turn-level memory* is too fine-grained, leading to fragmentary and incomplete context (*e.g.*, Turn-2 and Turn-5) and misses essential interaction turns (*e.g.*, Turn-4). *Session-level memory* is too coarse-grained, containing too much irrelevant information (*e.g.*, definition of the prosecutor's fallacy and reasons of the World War II), which distracts both the retrieval module and the response generation model. *Summary based methods* suffer from information loss that occurs during summarization. *Ours (segment-level memory)* can better capture topically coherent units in long conversations, striking a balance between including more relevant, coherent information while excluding irrelevant content. Bullseye ⊙ indicates the retrieved memory units at turn level or segment level. [0.xx]: similarity between target query and history content. Turn-level retrieval error: false negative, false positive.



(a) Response quality as a function of chunk size, given a total budget of 50 turns to retrieve as context.

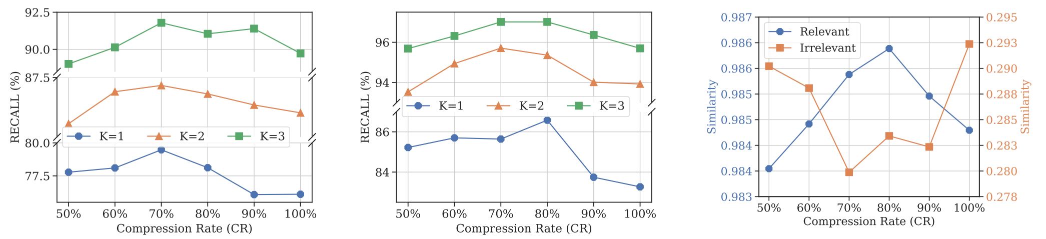
(b) Retrieval DCG obtained with different memory granularities using BM25 based retriever.

(c) Retrieval DCG obtained with different memory granularities using MPNet based retriever.

Figure 2: The impact of memory granularity on the response quality (a) and retrieval accuracy (b, c).

Long conversations are naturally composed of coherent discourse units. To capture this structure, we introduce a conversation segmentation model that partitions long-term conversations into topically coherent segments, constructing the memory bank at the segment level. During response generation, we directly concatenate the retrieved segment-level memory units as the context as in Yuan et al. [8] and Kim et al. [10], bypassing summarization to avoid the information loss that often occurs when converting dialogues into summaries [7].

Furthermore, inspired by the notion that natural language tends to be inherently redundant [15–17], we hypothesize that such redundancy can act as noise for retrieval systems, complicating the extraction of key information [18, 19]. Therefore, we propose removing such redundancy from memory units prior to retrieval by leveraging prompt compression methods such as LLMLingua-2 [17]. Figure 3 shows the results obtained with a BM25 based retriever and a MPNet based retriever [20] on *Long-MT-Bench+*. As demonstrated in Figure 3a and 3b, LLMLingua-2 consistently improves retrieval recall given different retrieval budgets K (*i.e.*, the number of retrieved segments) when the compression rate exceeds 50%. Figure 3c further illustrates that, after denoising, similarity between the query and relevant segments increases, while the similarity with irrelevant segments decreases.



(a) Retrieval recall v.s compression rate: $\frac{\text{\# tokens after compression}}{\text{\# tokens before compression}}$. K: number of retrieved segments. Retriever: BM25

(b) Retrieval recall v.s compression rate: $\frac{\text{\# tokens after compression}}{\text{\# tokens before compression}}$. K: number of retrieved segments. Retriever: MPNet

(c) Similarity between the query and different dialogue segments. Blue: relevant segments. Orange: irrelevant segments. Retriever: MPNet

Figure 3: Prompt compression method (LLMLingua-2) can serve as an effective denoising technique to enhance the memory retrieval system by: (a) improving the retrieval recall with varying context budget K; (b) benefit the retrieval system by increasing the similarity between the query and relevant segments while decreasing the similarity with irrelevant ones.

Our contributions can be summarized as follows:

- We systematically investigate the effects of memory granularity on retrieval augmented response generation in conversational agents. Our findings reveal that turn-level, session-level, and summarization-based approaches each face challenges in ensuring precise retrieval and providing a complete, relevant, and coherent context for generating accurate responses.
- We contend that the inherent redundancy in natural language can act as noise for retrieval systems. We demonstrate that prompt compression technique, LLMLingua-2, can serve as an effective denoising method to enhance memory retrieval performance.
- We present SECOM, a system that constructs memory bank at segment level by introducing a conversation SEgmentation model, while applying COMpression based denoising on memory units to enhance memory retrieval. The experimental results show that SECOM outperforms baselines on two long-term conversation benchmark LOCOMO and Long-MT-Bench+.

2 SeCom

2.1 Preliminary

Let $\mathcal{H} = \{\boldsymbol{c}_i\}_{i=1}^{C}$ represent the available conversation history between a user and an agent, which consists of C sessions. $\boldsymbol{c}_i = \{\boldsymbol{t}_j\}_{j=1}^{T_i}$ denotes the i-th session that is composed of T_i sequential user-agent interaction turns, with each turn $\boldsymbol{t}_j = (u_j, r_j)$ consisting of a user request u_j and the corresponding response from the agent r_j. Denote the base retrieval system as f_R and the response generation model as f_{LLM}. The research framework here can be defined as: (1) *Memory construction*: construct a memory bank $\mathcal{M}$ using conversation history $\mathcal{H}$. (2) *Memory retrieval*: given a target user request u^* and context budget N, retrieve N memory units $\{\boldsymbol{m}_n \in \mathcal{M}\}_{n=1}^{N} \leftarrow f_R(u^*, \mathcal{M}, N)$ that are relevant to user request u^*; (3) *Response generation*: take the retrieved N memory units in time order as the context and query the response generation model for response $r^* = f_{\text{LLM}}(u^*, \{\boldsymbol{m}_n\}_{n=1}^{N})$.

2.2 Conversation Segmentation

Given a conversation session $\boldsymbol{c}$, the conversation segmentation model $f_{\mathcal{I}}$ aims to split each session $\boldsymbol{c}_i$ into K_i topical segments $\{\boldsymbol{s}_k\}_{k=1}^{K_i}$. To achieve this, the segmentation model identifies *a set of segment indices* $\mathcal{I} = \{(p_k, q_k)\}_{k=1}^{K_i}$, where p_k and q_k represent the indexes of the first and last interaction turns for the k-th segment $\boldsymbol{s}_k$, with $p_k \leq q_k$, $p_{k+1} = q_k + 1$. This can be formulated as:

$$f_{\mathcal{I}}(\boldsymbol{c}_i) = \{\boldsymbol{s}_k\}_{k=1}^{K_i}, \text{where } \boldsymbol{s}_k = \{\boldsymbol{t}_{p_k}, \boldsymbol{t}_{p_k+1}, ..., \boldsymbol{t}_{q_k}\} \quad (1)$$

However, building a segmentation model for open-domain conversation is challenging, primarily due to the difficulty of acquiring large amounts of annotated data. As noted by Jiang et al. [21], the ambiguous nature of segmentation points complicates data collection, making the task difficult even for human annotators. Consequently, we employ GPT-4 as the conversation segmentation model $f_{\mathcal{I}}$. Figure 4 presents the detailed instruction used for conversation segmentation here.

2.3 Compression based Memory Denoising

Given a target user request u^* and context budget N, the memory retrieval system f_R retrieves N memory units $\{\boldsymbol{m}_n \in \mathcal{M}\}_{n=1}^{N}$ from the memory bank $\mathcal{M}$ as the context in response to the user request u^* . With the consideration that the inherent redundancy in natural language can act as noise for the retrieval system [18, 19], we denoise memory units by removing such redundancy via a prompt compression model f_{Comp} before retrieval:

$$\{\boldsymbol{m}_n \in \mathcal{M}\}_{n=1}^{N} \leftarrow f_R(u^*, f_{Comp}(\mathcal{M}), N). \quad (2)$$

Specifically, we use LLMLingua-2 [17] as the denoising function f_{Comp} here.

3 Experiments

Implementation Details We use `GPT-35-Turbo` in our main experiment. Details for the conversation segmentation are described in Appendix A.1. We use `LLMLingua-2` [17] with a compression rate of 75% and `xlm-roberta-large` [22] as the base model to denoise memory units. Following Alonso et al. [9], we apply MPNet (`multi-qa-mpnet-base-dot-v1`) [20] for memory retrieval.

We evaluate SeCom and other baseline methods on two benchmarks: (i) *LOCOMO* [7], which is the longest conversation dataset to date, with an average of 300 turns per sample. (ii) *Long-MT-Bench+*, which is reconstructed from *MT-Bench+* [23]. More details are provided in Appendix A.5. For evaluation metrics, we use the conventional *BLEU* [24], *ROUGE* [25], and *BERTScore* [26]. We also employ *GPT4Score* [23] for more accurate evaluation. The evaluation details are in Appendix A.3.

Baselines (1) *Turn-Level*, which treats each user-agent interaction as an individual memory unit. (2) *Session-Level*, which uses each entire conversation session as a memory unit. (3) *Zero History*, which does not use any conversation history. (4) *Full History*, which concatenates all prior conversation history as the context for response generation. (5) *SumMem* [27], which summarizes past dialogues and uses these summaries as context for response generation. (6) *RecurSum* [6], which recursively updates summary using current session and previous summaries, and takes the updated summary as the context. (7) *ConditionMem* [8], which generates summaries and knowledge as memory records, then retrieves relevant memory records as the context. (8) *MemoChat* [23], which operates memories at segment level, but focuses on tuning LLMs for both memory construction and retrieval.

Main Results As shown in Table 1, SeCom*, which constructs memory bank at segment level, outperforms turn-level and session-level approaches*, exhibiting a significant performance advantage, particularly on the long-conversation benchmark *LOCOMO*. We attribute this to the following reason: As discussed in Section 1, turn-level memory units are often fragmented and may not explicitly include or relate to keywords mentioned in the target user request. On the other hand, session-level memory units contain excessive irrelevant information. Both of these scenarios make the retrieval performance sensitive to the capability of the deployed retrieval system. However, topical segments in SeCom can strike a balance between including more relevant, coherent information while excluding irrelevant content, thus leading to more robust and superior retrieval performance. Table 1 also reveals that *summary based methods, such as SumMem and RecurSum fall behind turn-level or session-level baselines*. Our case study in Appendix A.7 suggests that this is likely due to the loss of details essential for accurate question answering when converting dialogues into summaries [7].

Table 1: Performance comparison on *LOCOMO* and *Long-MT-Bench+*. The retrieval budget is set to 4k tokens ($\sim$ 55 turns) on *LOCOMO* and 1k tokens ($\sim$ 3 turns) on *Long-MT-Bench+*.

Methods	**QA Performance**						**Context Length**	
	GPT4Score	BLEU	Rouge1	Rouge2	RougeL	BERTScore	# Turns	# Tokens
				LOCOMO				
Zero History	24.86	1.94	17.36	3.72	13.24	85.83	0.00	0
Full History	54.15	6.26	27.20	12.07	22.39	88.06	210.34	13,330
Turn-Level	57.99	6.07	26.61	11.38	21.60	88.01	54.77	3,288
Session-Level	51.18	5.22	24.23	9.33	19.51	87.45	53.88	3,471
SumMem	53.87	2.87	20.71	6.66	16.25	86.88	-	4,108
RecurSum	56.25	2.22	20.04	8.36	16.25	86.47	-	400
ConditionMem	<u>65.92</u>	3.41	22.28	7.86	17.54	87.23	-	3,563
MemoChat	65.10	<u>6.76</u>	<u>28.54</u>	<u>12.93</u>	<u>23.65</u>	<u>88.13</u>	-	1,159
SECOM	**69.33**	**7.19**	**29.58**	**13.74**	**24.38**	**88.60**	55.51	3,716
				Long-MT-Bench+				
Zero History	49.73	4.38	18.69	6.98	13.94	84.22	0.00	0
Full History	63.85	7.51	26.54	12.87	20.76	85.90	65.45	19,287
Turn-Level	84.91	12.09	<u>34.31</u>	<u>19.08</u>	**27.82**	86.49	3.00	909
Session-Level	73.38	8.89	29.34	14.30	22.79	86.61	13.43	3,680
SumMem	63.42	7.84	25.48	10.61	18.66	85.70	-	1,651
RecurSum	62.96	7.17	22.53	9.42	16.97	84.90	-	567
ConditionMem	63.55	7.82	26.18	11.40	19.56	86.10	-	1,085
MemoChat	<u>85.14</u>	<u>12.66</u>	33.84	19.01	26.87	<u>87.21</u>	-	1,615
SECOM	**88.81**	**13.80**	**34.63**	**19.21**	<u>27.64</u>	**87.72**	2.77	820

Effectiveness of the Conversation Segmentation Model We evaluate the LLM-based segmentation module in three widely-used dialogue segmentation datasets: DialSeg711 [28], TIAGE [29], and SuperDialSeg [21]. Table 3 in Appendix A.1 illustrates the effectiveness of our segmentation model, showcasing a significant performance advantage over state-of-the-art unsupervised and transfer-learning based baselines. This advantage underscores the model's suitability for segmenting long-term, open-domain conversation where annotated data are difficult to obtain [21].

Ablation Study on Compression based Memory Denoising As shown in Table 2, removing the compression based memory denoising mechanism will result in a performance drop up to 9.46 points of GPT4Score on LOCOMO, highlighting the critical role of this denoising mechanism: by improving the retrieval system (Figure 3b), it finally enhances the overall effectiveness of the system.

Table 2: Ablation study on compression based memory denoising with a compression rate of 75%.

Methods	**LOCOMO**				**Long-MT-Bench+**			
	GPT4Score	BLEU	Rouge2	BERTScore	GPT4Score	BLEU	Rouge2	BERTScore
SECOM	**69.33**	**7.19**	**13.74**	**88.60**	**88.81**	**13.80**	**19.21**	**87.72**
− Denoise	59.87	6.49	12.11	88.16	87.51	12.94	18.73	87.44

4 Conclusion

In this paper, we systematically investigate the impact of memory granularity on retrieval-augmented response generation for long-term conversational agents, revealing the limitations of turn-level and session-level memory granularities, as well as summarization-based methods. To overcome these challenges, we introduce SECOM, a novel memory management system that constructs memory bank at the segment-level and employs compression-based denoising techniques to enhance retrieval performance. The experimental results underscore the effectiveness of SECOM.

References

[1] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In International Conference on Learning Representations, 2019.

[2] Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. Aligning large language models through synthetic feedback. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 13677–13700, 2023.

[3] Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1714–1729, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.96. URL `https://aclanthology.org/2023.acl-long.96`.

[4] Nuo Chen, Hongguang Li, Juhua Huang, Baoyuan Wang, and Jia Li. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. arXiv preprint arXiv:2402.11975, 2024.

[5] LangChain Team. Conversation buffer. `https://python.langchain.com/v0.1/docs/modules/memory/types/buffer/`, 2023.

[6] Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. Recursively summarizing enables long-term dialogue memory in large language models. arXiv preprint arXiv:2308.15022, 2023.

[7] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. arXiv preprint arXiv:2402.17753, 2024.

[8] Ruifeng Yuan, Shichao Sun, Zili Wang, Ziqiang Cao, and Wenjie Li. Evolving large language model assistant with long-term conditional memory. arXiv preprint arXiv:2312.17257, 2023.

[9] Nick Alonso, Tomás Figliolia, Anthony Ndirango, and Beren Millidge. Toward conversational agents with context and time sensitive long-term memory. arXiv preprint arXiv:2406.00057, 2024.

[10] Seo Hyun Kim, Kai Tzu-iunn Ong, Taeyoon Kwon, Namyoung Kim, Keummin Ka, SeongHyeon Bae, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. Theanine: Revisiting memory management in long-term conversations with timeline-augmented response generation. arXiv preprint arXiv:2406.10996, 2024.

[11] Jing Xu, Arthur Szlam, and Jason Weston. 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, 2022.

[12] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. Proceedings of the AAAI Conference on Artificial Intelligence, 38(17):19724–19731, Mar. 2024. doi: 10.1609/aaai.v38i17.29946. URL `https://ojs.aaai.org/index.php/AAAI/article/view/29946`.

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474, 2020.

[14] Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 2301–2305, 2024.

[15] Claude E Shannon. Prediction and entropy of printed english. Bell system technical journal, 30(1):50–64, 1951.

[16] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 13358–13376, 2023.

[17] Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. arXiv preprint arXiv:2403.12968, 2024.

[18] David Grangier, Alessandro Vinciarelli, and Hervé Bourlard. Information retrieval on noisy text. In IDIAP COMMUNICATION, 2003. URL `https://api.semanticscholar.org/CorpusID:3249973`.

[19] Xueguang Ma, Minghan Li, Kai Sun, Ji Xin, and Jimmy Lin. Simple and effective unsupervised redundancy elimination to compress dense vectors for passage retrieval. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2854–2859, 2021.

[20] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. Advances in neural information processing systems, 33:16857–16867, 2020.

[21] Junfeng Jiang, Chengzhang Dong, Sadao Kurohashi, and Akiko Aizawa. Superdialseg: A large-scale dataset for supervised dialogue segmentation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4086–4101, 2023.

[22] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL `https://aclanthology.org/2020.acl-main.747`.

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

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

[25] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81, 2004.

[26] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In International Conference on Learning Representations, 2020.

[27] LangChain Team. Conversation summary memory. `https://python.langchain.com/v0.1/docs/modules/memory/types/summary/`, 2023.

[28] Yi Xu, Hai Zhao, and Zhuosheng Zhang. Topic-aware multi-turn dialogue modeling. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 14176–14184, 2021.

[29] Huiyuan Xie, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, and Ann Copestake. Tiage: A benchmark for topic-shift aware dialog modeling. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1684–1690, 2021.

[30] André V Duarte, João Marques, Miguel Graça, Miguel Freire, Lei Li, and Arlindo L Oliveira. Lumberchunker: Long-form narrative document segmentation. arXiv preprint arXiv:2406.17526, 2024.

[31] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS), 20(4):422–446, 2002.

[32] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In International Conference on Machine Learning, pages 31210–31227. PMLR, 2023.

[33] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173, 2024.

A Appendix

A.1 Details of the Conversation Segmentation Model

We use GPT-4-0125 as the backbone LLM for conversation segmentation. The segmentation prompt is provided in Figure 4. It instructs the segmentation model to generate all segmentation indices at once, avoiding the iterative segmentation process used in LumberChunker [30], which can lead to unacceptable latency. We specify that the output should be in **JSONL** format to facilitate subsequent processing. Table 3 presents the segmentation results. Our LLM-based segmentation model demonstrates significantly better performance compared to state-of-the-art unsupervised and transfer-learning baselines, highlighting its suitability for segmenting long-term, open-domain conversations where annotated data are scarce [21].

Instruction Part of the Segmentation Prompt (Zero-Shot).

```
# Instruction
## Context
- **Goal**: Your task is to segment a multi-turn conversation between a
user and a chatbot into topically coherent units based on semantics.
Successive user-bot exchanges with the same topic should be grouped
into the same segmentation unit, and new segmentation units should
be created when topic shifts.
- **Data**: The input data is a series of user-bot exchanges separated
by "\n\n". Each exchange consists of a single-turn conversation between
the user and the chatbot, started with "[Exchange (Exchange Number)]: ".
### Output Format
- Output the segmentation results in **JSONL (JSON Lines)** format.
Each dictionary represents a segment, consisting of one or more
user-bot exchanges on the same topic.
Each dictionary should include the following keys:
  - **segment_id**: The index of this segment, starting from 0.
  - **start_exchange_number**: The number of the **first** user-bot
  exchange in this segment.
  - **end_exchange_number**: The number of the **last**
  user-bot exchange in this segment.
  - **num_exchanges**: An integer indicating the number of
  user-bot exchanges in this segment, calculated as:
  **end_exchange_number** - **start_exchange_number** + 1.
Here is an example of the expected output:
‘‘‘
<segmentation>
{"segment_id": 0, "start_exchange_number": 0,
"end_exchange_number": 5, "num_exchanges": 6}
{"segment_id": 1, "start_exchange_number": 6,
"end_exchange_number": 8, "num_exchanges": 3}
...
</segmentation>
‘‘‘
# Data
{{text_to_be_segmented}}
# Question
## Please generate the segmentation result from the input data that
meets the following requirements:
- **No Missing Exchanges**:  Ensure that the exchange numbers cover
all exchanges in the given conversation without omission.
- **No Overlapping Exchanges**: Ensure that successive segments have
no overlap in exchanges.
- **Accurate Counting**:  The sum of **num_exchanges**
across all segments should equal the total number of user-bot exchanges.
- Provide your segmentation result between the tags:
<segmentation></segmentation>.
# Output
Now, provide the segmentation result based on the instructions above.
```

Figure 4: Prompt for GPT-4 segmentation (zero-shot).

Table 3: Segmentation performances on three datasets. Numbers of baselines are reported in Jiang et al. [21]. The best performance is highlighted in **bold**, and the second best is highlighted by underline. For *DialSeg711*, the transfer-learning baselines are trained on *TIAGE* [29], as it yields better performance than training on *SuperDialSeg* [21].

Methods	Dialseg711				SuperDialSeg				TIAGE			
	Pk↓	WD↓	F1↑	Score↑	Pk↓	WD↓	F1↑	Score↑	Pk↓	WD↓	F1↑	Score↑
Unsupervised Baselines												
BayesSeg	0.306	0.350	0.556	0.614	<u>0.433</u>	0.593	<u>0.438</u>	0.463	0.486	0.571	0.366	0.419
TextTiling	0.470	0.493	0.245	0.382	0.441	0.453	0.388	<u>0.471</u>	0.469	0.488	0.204	0.363
GraphSeg	0.412	0.442	0.392	0.483	0.450	0.454	0.249	0.398	0.496	0.515	0.238	0.366
TextTiling+Glove	0.399	0.438	0.436	0.509	0.519	0.524	0.353	0.416	0.486	0.511	0.236	0.369
TextTiling+[CLS]	0.419	0.473	0.351	0.453	0.493	0.523	0.277	0.385	0.521	0.556	0.218	0.340
TextTiling+NSP	0.347	0.360	0.347	0.497	0.512	0.521	0.208	0.346	0.425	0.439	0.285	0.426
GreedySeg	0.381	0.410	0.445	0.525	0.490	0.494	0.365	0.437	0.490	0.506	0.181	0.341
CSM	0.278	0.302	0.610	0.660	0.462	0.467	0.381	0.458	<u>0.400</u>	0.420	<u>0.427</u>	<u>0.509</u>
Transfer-learning Based Baselines												
Training Set	**Train on TIAGE**				**Train on TIAGE**				**Train on SuperDialSeg**			
TextSeg$_{dial}$	0.476	0.491	0.182	0.349	0.552	0.570	0.199	0.319	0.489	0.508	0.266	0.384
BERT	0.441	0.411	0.005	0.297	0.511	0.513	0.043	0.266	0.492	0.526	0.226	0.359
RoBERTa	<u>0.197</u>	<u>0.210</u>	<u>0.650</u>	<u>0.723</u>	0.434	<u>0.436</u>	0.276	0.420	0.401	<u>0.418</u>	0.373	0.482
LLM-based Segmentation Model (Zero-Shot)												
Ours	**0.093**	**0.103**	**0.888**	**0.895**	**0.277**	**0.289**	**0.758**	**0.738**	**0.363**	**0.401**	**0.596**	**0.607**

Table 4: Comparison between our method and *MemoChat* from multiple aspects on *Long-MT-Bench+*. “# In. Token", “# Out. Token" and “Latency" report the number of input / output token and the latency per question, including memory construction, memory retrieval and reponse generation.

Methods	# In. Token	# Out. Token	Latency (s)	GPT Score
Session-Level	3,642	102	2.17	73.38
MemoChat	7,233	229	5.60	85.14
Ours	1,722	135	2.61	**88.81**

A.2 Additional Cost Analysis

Table 4 compares the overall costs involved in memory construction, memory retrieval, and response generation across different methods. The results demonstrate that our method significantly enhances performance compared to the baseline while only slightly increasing computational overhead, and it outperforms the MemoChat method in both efficiency and effectiveness.

A.3 Prompt for GPT-4 Evaluation

We use the same evaluation prompts as MemoChat [23]. The LLM-powered evaluation consists of single-sample scoring (GPT4Score) and pair-wise comparison. The evaluation prompts are displayed in Figure 5. For pair-wise comparison, we alternate the order of the responses and conduct a second comparison for each pair to minimize position bias.

A.4 Evaluation Results on the Official QA Pairs of LOCOMO

As *LOCOMO* [7] released a subset containing QA pairs recently. To ensure reproducibility, we evaluate our method on these official QA pairs. Table 5 presents the evaluation results. The superiority of our SeCom is also evident on these QA pairs, demonstrating its superior effectiveness and robustness.

Single-Sample Score

```
You are an impartial judge. You will be shown Related
Conversation History, User Question and Bot Response.
‘‘‘\nRelated Conversation History\nRCH\_0\n‘‘‘
‘‘‘\nUser Question\nUQ\_1\n‘‘‘
‘‘‘\nBot Response\nBR\_2\n‘‘‘
Please evaluate whether Bot Response is faithful to the content of
Related Conversation History to answer User Question.
Begin your evaluation by providing a short explanation,
then you must rate Bot Response on an integer rating of 1 to
100
by strictly following this format:
<rating>an integer rating of 1 to 100</rating>.
```

Pair-Wise Comparison

```
You are an impartial judge. You will be shown
Related Conversation History, User Question and Bot Response.
‘‘‘\nRelated Conversation History\nRCH_0\n‘‘‘
‘‘‘\nUser Question\nUQ_1\n‘‘‘
‘‘‘\nBot Response A\nBR_2\n‘‘‘
‘‘‘\nBot Response B\nBR_3\n‘‘‘
Please evaluate whether Bot Response is faithful to the content of
Related Conversation History to answer User Question.
Begin your evaluation by
providing a short explanation,
then you must choose the better bot response by giving
either A or B.
If the two responses are the same, you can choose NONE:
<chosen>A (or B or NONE)</chosen>.
```

Figure 5: Prompt used in GPT-4 evaluation, following Lu et al. [23].

Table 5: Performance comparison on the official question-answer pairs of *LOCOMO*. All settings are the same as in Table 1.

Methods	QA Performance						Context Length	
	GPT4Score	BLEU	Rouge1	Rouge2	RougeL	BERTScore	# Turns	# Tokens
Full History	66.28	7.51	28.73	14.07	27.90	87.82	293	18,655
MemoChat	75.77	11.28	32.91	18.82	29.78	87.98	-	1,159
Turn-Level	81.52	11.91	36.00	19.59	34.99	**88.64**	55.00	3,026
Session-Level	74.20	10.95	29.92	14.64	29.27	87.96	54.48	3,442
SeCom	**84.21**	**12.80**	**36.70**	**19.90**	**35.61**	88.59	56.49	3,565

A.5 Details of Dataset Construction

(i) *LOCOMO* [7]: this dataset contains the longest conversations to date, with an average of more than 9K tokens per sample. Since *LOCOMO* does not release the corresponding question-answer pairs when we conduct our experiment, we prompt GPT-4 to generate QA pairs for each session as in Alonso et al. [9]. We also conduct evaluation on the recently released official QA pairs in Appendix A.4.

(ii) *Long-MT-Bench+*: *Long-MT-Bench+* is reconstructed from the *MT-Bench+* [23] dataset. In *MT-Bench+*, human experts are invited to expand the original questions and create long-range questions as test samples. However, there are two drawbacks when using this dataset to evaluate the memory mechanism of conversational agents: (1) the number of test set QA pairs is relatively small, with only 54 human-written long-range questions; and (2) the conversation length is not sufficiently long,

Table 6: Statistics of the *MT-Bench+* and the constructed *Long-MT-Bench+* datasets. The notation *"# Item"* represents the average number of the corresponding item per conversation.

Datasets	# QA. Pairs	# Session	# Round	# Token
MT-Bench+	1	1	13.33	3,929
Long-MT-Bench+	26.09	4.91	65.45	19,287

with each conversation containing an average of 13.3 dialogue turns and a maximum of 16 turns. In contrast, the conversation in *LOCOMO* has an average of 300 turns and 9K tokens. To address these limitations, we use the human-written questions as few-shot examples and prompt GPT-4 to generate long-range test questions for each dialogue topic, thereby expanding the QA pairs in the test set. Additionally, following Yuan et al. [8], we merge five consecutive sessions into one to create longer dialogues that are more suitable for evaluating memory in long-term conversations. We refer to this reconstructed dataset as *Long-MT-Bench+* and present its statistics in Table 6.

A.6 Details of Retrieval Performance Measurement

We measure the retrieval performance in terms of the discounted cumulative gain (DCG) metric [31]:

$$DCG = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}, \tag{3}$$

where rel_i denotes the relevance score of the retrieved user-agent turn ranked at position i, and p represents the total number of retrieved turns. Note that in the *Long-MT-Bench+* dataset, answering a single question often requires referring to several consecutive turns. Therefore, we distribute the relevance score evenly across these relevant turns and set the relevance score of irrelevant turns to zero. For instance, assume that the ground truth reference turn set for question q is $\mathcal{R}(q) = \{r_{k+j}\}_{j=1}^{N}$, which is provided by the dataset. In this case, the relevance score for each turn is set as follows:

$$rel_i = \begin{cases} 0 & i < k+1 \\ \frac{1}{N} & k+1 \leq i \leq k+N \\ 0 & i > k+N \end{cases}.$$

This approach allows us to evaluate retrieval performance at different granularity.

A.7 Case Study

To further demonstrate the advantages of our method, we conduct a qualitative evaluation. Figure 6 presents a specific case comparing the segment-level memory with the turn-level memory. It demonstrates that using turn-level memory units fails to address the user's request. We attribute this to the fragmentation of user-agent turns, and the critical turns may not explicitly contain or relate to the keywords in the user's request.

Similarly, using session-level memory units is also sub-optimal, as illustrated in Figure 7. This issue arises because a session often includes multiple topics, introducing a significant amount of irrelevant information that hampers effective retrieval. The irrelevant information also distracts the LLM, as noted in previous studies [32, 33].

We also conduct a case study to compare our method with two recent, powerful memory management techniques: *RecurSum* [6] and *ConditionMem* [8], as shown in Figure 8 and Figure 9. The results indicate that the summarization process in these methods often omits detailed information that is essential for accurately answering the user's request.

Conversation History:

[human]: Craft an intriguing opening paragraph for a fictional short story. The story should involve a character who wakes up one morning to find that they can time travel.
...(Human-Bot Dialogue Turns)... (Topic: Time-Travel Fiction)
[human]: Please describe the concept of machine learning. Could you elaborate on the differences between supervised, unsupervised, and reinforcement learning? Provide real-world examples of each.
...(Human-Bot Dialogue Turns)... (Topic: Machine learning Concepts and Types)
[human]: Discuss antitrust laws and their impact on market competition. Compare the antitrust laws in US and China along with some case studies
...(Human-Bot Dialogue Turns)... (Topic: Antitrust Laws and Market Competition)
[human]: The vertices of a triangle are at points (0, 0), (-1, 1), and (3, 3). What is the area of the triangle?
...(Human-Bot Dialogue Turns)... (Topic: Area and Circumcircle of Triangle)
...(Human-Bot Dialogue Interaction on Other Topics)...

Question:

What title reflects the thematic essence of the Doctor's experience with time travel as described in the story?

Ground-Truth Answer:

Awakening to Eternity: The Doctor's New Dawn

Retrieved Memory (Turn-Level):

[human]: Describe the process and write out the balanced chemical equation for the reaction that occurs when solid calcium carbonate reacts with hydrochloric acid to form aqueous calcium chloride, carbon dioxide, and water. What type of reaction is this, and what observations might indicate that the reaction is taking place?
[bot]: When solid calcium carbonate ($CaCO_3$) reacts with hydrochloric acid (HCl), the products formed are aqueous calcium chloride ($CaCl_2$), carbon dioxide (CO_2), and water (H_2O). The balanced chemical equation for this reaction is: ... (Topic: Calcium Carbonate and Hydrochloric Acid Reaction)
[human]: Write a title for your story.
[bot]: "Awakening to Eternity: The Doctor's New Dawn". (Topic: Time-Travel Fiction)
[human]: Analyze the following customer reviews from different sources for three different smartphones - the latest iPhone, Samsung Galaxy, and Google Pixel - and provide an overall rating for each phone on a scale of 1 to 10. Consider the following complex and contradictory reviews: - TechRadar's review of the latest iPhone: The new iPhone is a stunning triumph of engineering that sets a new bar for smartphone performance and camera quality. However, the incremental design and high price mean it lacks the 'wow' factor of previous iPhones. Still, its power and intelligence are unrivaled. - CNET's review of the latest Samsung Galaxy: ... Return the answer as a JSON object with the overall ratings for each phone out of 10, to one decimal place.
[bot]: “‘json "iPhone": 8.5, "Samsung Galaxy": 8.0, "Google Pixel": 7.5 “‘
(Topic: Smartphone Review Ratings Analysis)

Answer (Turn-Level Retrieval):

As an AI language model, I do not have access to the story you mentioned earlier. Please provide me with the story or the necessary details so that I can assist you better.
(The next part of this figure 6 will be continued on the next page.)

Retrieved Memory (Ours):

[human]: Craft an intriguing opening paragraph for a fictional short story. The story should involve a character who wakes up one morning to find that they can time travel.
[bot]: In the hushed pre-dawn hours, when the world still slumbers under the cloak of night, Alex found themselves caught in the fragile boundary between sleep and wakefulness. ...
[human]: Summarize the story with three bullet points using only nouns and adjectives, without verbs.
[bot]: - Alex, pre-dawn awakening, time-travel discovery - Disorientation, vertigo, reality's malleability - Sunlight, new day, astonishing truth
[human]: Is there any famous fictional character that can time travel? If so, can you rewrite your story with him or her?
[bot]: - Doctor Who, a renowned time-traveling character from British television, is a perfect fit for this scenario. In the silent embrace of early morning, where shadows merge with the fading night, the Doctor found themselves in the thin veil separating dreams from consciousness. It was here, amidst the soft whispers of time, that an unusual sensation enveloped them—a sense of time's fabric slipping through their fingers like grains of sand. As dawn's light broke through the darkness, painting the room in hues of gold and amber, the Doctor's eyes opened to what appeared to be an ordinary day. However, upon standing, a wave of dizziness overwhelmed them, the surroundings blurred, and in that moment of confusion, the Doctor realized a profound truth: they had not only awakened to a new dawn but had also stumbled upon the incredible ability to navigate the currents of time at will.
[human]: Write a title for your story.
[bot]: "Awakening to Eternity: The Doctor's New Dawn" (Topic: Time-Travel Fiction)

Answer (Ours):

The title "Awakening to Eternity: The Doctor's New Dawn" reflects the thematic essence of the Doctor's experience with time travel as described in the story. It captures the idea of a new beginning, a sense of eternity, and the awakening of a new power that will change the course of the Doctor's life forever.

Figure 6: Case study comparing our SeCom with the *turn-level* memory baseline. When constructing the memory bank at the *turn-level*, the agent fails to provide a satisfactory response due to fragmentation: the retrieved dialogue turns can not capture a complete human-agent interaction related to the query, missing critical information highlighted in brown. The dialogue topic is emphasized in blue. Some context is omitted (replaced by ...) due to length constraints.

Conversation History:

[human]: Explain what's base rate fallacy and list five specific examples of ...
... (Human-Bot Dialogue Turns) ... (Topic: Base Rate Fallacy In Politics)
[human]: How many integers are in the solution of the inequality $|x + 5| < 10$?
... (Human-Bot Dialogue Turns) ... (Topic: Number of Integers in Inequalities)
[human]: How have the Alps and Rhine River influenced settlement and agriculture in Western Europe?
... (Human-Bot Dialogue Turns) ... (Topic: Alps and Rhine River Impact)
[human]: In a survey conducted at a local high school, preferences for a new school color were measured: 58% of students liked the color blue, 45% preferred green, and 22% liked both colors. If we randomly pick a student from the school, what's the probability that they would like neither blue nor green?
... (Human-Bot Dialogue Turns) ... (Topic: Probability of Liking Colors in Survey)
[human]: Imagine you are participating in a race with a group of people. If you have just overtaken the second person, what's your current position? Where is the person you just overtook?
... (Human-Bot Dialogue Turns) ... (Topic: Overtaking Positions in Race)

Question:

[human]: In the inequality question, how many positive integers are in the solution of the second inequality?

Ground-Truth Answer:

In the second inequality, ($|x + 10| < 5$), the solution range is ($-15 < x < -5$). Within this range, there are no positive integers, as all values of (x) that satisfy the inequality are negative.

Retrieved Memory (Session-Level):

[human]: Explain what's base rate fallacy and list five specific examples of ...
... (Human-Bot Dialogue Turns) ... (Topic: Base Rate Fallacy In Politics)
[human]: How many integers are in the solution of the inequality $|x + 5| < 10$?
[bot]: To solve the inequality $|x + 5| < 10$, ...
[human]: What about $|x + 10| < 5$?
[bot]: To solve the inequality $|x + 10| < 5$, we ... ("overwhelmed" by other irrelevant, redundant content)
... (Human-Bot Dialogue Turns) ... (Topic: Number of Integers in Inequalities)
[human]: How have the Alps and Rhine River influenced settlement and agriculture in Western Europe?
... (Human-Bot Dialogue Turns) ... (Topic: Alps and Rhine River Impact)
[human]: In a survey conducted at a local high school, ...
... (Human-Bot Dialogue Turns) ... (Topic: Probability of Liking Colors in Survey)

Answer (Session-Level Retrieval):

I apologize for the confusion earlier. To answer your question, the second inequality was not provided in our previous conversation. If you could provide the inequality, I would be happy to help you find the solution and the number of positive integers in the solution set.
(The next part of this figure 7 will be continued on the next page.)

Retrieved Memory (Ours):

[human]: How many integers are in the solution of the inequality $|x + 5| < 10$?
[bot]: To solve the inequality $|x + 5| < 10$, we break it into two cases based on the definition of absolute value: 1. $x + 5 < 10$ 2. $-(x + 5) < 10$, which simplifies to $x + 5 > -10$ Solving each case: 1. $x < 5$ 2. $x > -15$ Combining these, we get $-15 < x < 5$. This inequality means that x can be any number greater than -15 and less than 5. To find how many integers satisfy this condition, we list them: $-14, -13, -12, -11, -10, -9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4$. There are 19 integers that satisfy the inequality $|x + 5| < 10$.
[human]: What about $|x + 10| < 5$?
[bot]: To solve the inequality $|x + 10| < 5$, we again consider two cases: 1. $x + 10 < 5$ 2. $-(x + 10) < 5$, which simplifies to $x + 10 > -5$ Solving each case: 1. $x < -5$ 2. $x > -15$ Combining these, we get $-15 < x < -5$. This inequality means that x can be any number greater than -15 and less than -5. To find how many integers satisfy this condition, we list them: $-14, -13, -12, -11, -10, -9, -8, -7, -6$. There are 9 integers that satisfy the inequality $|x + 10| < 5$.
[human]: What about $|x + 10| < 10$?
[bot]: To solve the inequality $|x + 10| < 10$, we consider two cases: 1. $x + 10 < 10$ 2. $-(x + 10) < 10$, which simplifies to $x + 10 > -10$ Solving each case: 1. $x < 0$ 2. $x > -20$ Combining these, we get $-20 < x < 0$. This inequality means that x can be any number greater than -20 and less than 0. To find how many integers satisfy this condition, we list them: $-19, -18, -17, -16, -15, -14, -13, -12, -11, -10, -9, -8, -7, -6, -5, -4, -3, -2, -1$. There are 19 integers that satisfy the inequality $|x + 10| < 10$. (Topic: Number of Integers in Inequalities)

Answer (Ours):

To find the number of positive integers that satisfy the inequality ($|x+10| < 5$), we need to look for positive integers in the solution set ($-15 < x < -5$). Since there are no positive integers in this interval, the answer is 0.

Figure 7: Case study comparing our SECOM with the *session-level* memory baseline. When constructing the memory bank at the *session-level*, the agent fails to provide a satisfactory response due to the inclusion of excessive irrelevant information within a session. This distracts the generator from focusing on the truly useful information highlighted in brown. The dialogue topic is emphasized in blue. Some context is omitted (replaced by ...) due to length constraints.

Conversation History:

[human]: Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage? ... (Human-Bot Dialogue Turns)... (Topic: Photosynthetic Energy Production)
[human]: Please assume the role of an English translator, tasked with correcting and enhancing spelling and language. Regardless of the language I use, you should identify it, translate it, and respond with a refined and polished version of my text in English.
... (Human-Bot Dialogue Turns)... (Topic: Language Translation and Enhancement)
[human]: Suggest five award-winning documentary films with brief background descriptions for aspiring filmmakers to study.
[bot]: ... 5. Än Inconvenient Trutḧ(2006) - Directed by Davis Guggenheim and featuring former United States Vice President Al Gore, this documentary aims to educate the public about global warming. It won two Academy Awards, including Best Documentary Feature. The film is notable for its straightforward yet impactful presentation of scientific data, making complex information accessible and engaging, a valuable lesson for filmmakers looking to tackle environmental or scientific subjects.
... (Human-Bot Dialogue Turns)... (Topic: Documentary Films Recommendation)
[human]: Given the following records of stock prices, extract the highest and lowest closing prices for each month in the year 2022. Return the results as a CSV string, with one line allocated for each month. Date,Open,High,Low,Close,Volume ... ... (Human-Bot Dialogue Turns)... (Topic: Stock Prices Analysis)
[human]: The city of Vega intends to build a bridge that will span the Vegona River, covering a distance of 1.8 kilometers. The proposed location falls within a seismically active area that has experienced several high-magnitude earthquakes. Given these circumstances, what would be the best approach to constructing the bridge? ... (Human-Bot Dialogue Turns)... (Topic: Seismic Bridge Construction)

Question:

Which documentary won two Academy Awards and what lesson does it offer to filmmakers?

Ground-Truth Answer:

“An Inconvenient Truth" (2006) won two Academy Awards. It offers the lesson of making complex information accessible and engaging, a valuable lesson for filmmakers looking to tackle environmental or scientific subjects.

Memory generated by RecurSum [6]:

The dialogue provided detailed information on various topics, including the stages of photosynthesis, energy production estimates through photosynthesis for different organisms, translations and enhancements of phrases in different languages, recommendations for documentary films, a pitch for a film about overcoming adversity, analysis of stock prices, and considerations for constructing a bridge in a seismically active area. Here’s a summary of the key points: ...
Documentary Film Recommendations: Suggested five award-winning documentary films for aspiring filmmakers to study, with brief background descriptions for each. The response included:
- "March of the Penguins" (2005) - Emperor penguins’ journey in Antarctica; noted for its cinematography and portrayal of the penguins’ life cycle.
- "Citizenfour" (2014) - Edward Snowden and NSA spying scandal; known for its real-time exploration of Snowden’s revelations.
- "20 Feet from Stardom" (2013) - Stories of backup singers behind musical legends; celebrates unsung heroes in the music industry.
- "Man on Wire" (2008) - Philippe Petit’s 1974 high-wire walk between the Twin Towers; combines personal footage, interviews, and re-enactments.
- "An Inconvenient Truth" (2006) - Al Gore’s documentary on global warming; known for its presentation of scientific data.

Answer (RSum):

I’m sorry, but based on the updated context you provided, I don’t see any previous discussion related to the question you asked. However, to answer your question, the documentary that won two Academy Awards is "Free Solo," which won for Best Documentary Feature and Best Achievement in Film Editing in 2019. .... The film follows climber Alex Honnold as he attempts to free solo climb El Capitan in Yosemite National Park, and the filmmakers had to take risks themselves in order to capture the footage and tell the story.
(The next part of this figure 8 will be continued on the next page.)

Retrieved Memory (Ours):

[human]: Suggest five award-winning documentary films with brief background descriptions for aspiring filmmakers to study.
[bot]: 1. **"March of the Penguins" (2005)** - Directed by Luc Jacquet, this French feature-length nature documentary delves into the annual journey of Emperor penguins across the Antarctic. It won the Academy Award for Best Documentary Feature. The film is celebrated for its stunning cinematography and its detailed, heartfelt portrayal of the penguins' life cycle, making it a must-study for filmmakers interested in wildlife documentaries. 2. **"Citizenfour" (2014)** ... 5. **"An Inconvenient Truth" (2006)** - Directed by Davis Guggenheim and featuring former United States Vice President Al Gore, this documentary aims to educate the public about global warming. It won two Academy Awards, including Best Documentary Feature. The film is notable for its straightforward yet impactful presentation of scientific data, making complex information accessible and engaging, a valuable lesson for filmmakers looking to tackle environmental or scientific subjects.
[human]: With the spirit in the first film, craft a succinct and persuasive pitch for a film about overcoming adversity.
[bot]: Drawing inspiration from the resilience and majestic journey of Emperor penguins in "March of the Penguins," our documentary, "Unyielding Spirits: The Human Odyssey," will explore the profound stories of individuals around the globe who have faced and overcome staggering adversities. Through breathtaking cinematography and intimate storytelling, we ... (Topic: Documentary Films Recommendation)

Answer (Ours):

“An Inconvenient Truth" won two Academy Awards, including Best Documentary Feature. The film offers a valuable lesson to filmmakers looking to tackle environmental or scientific subjects, as it presents complex information in a straightforward yet impactful way, making it accessible and engaging to a wider audience.

Figure 8: Case study comparing our SeCom with RecurSum [6]. When using RecurSum, the agent fails to provide a satisfactory answer due to information loss during the summarization process. Summarization omits essential details needed to answer the user's new request, as highlighted in brown. The dialogue topic is emphasized in blue. Some context is omitted (replaced by ...) due to length constraints.

Conversation History:

[human]: Photosynthesis is a vital process for life on Earth. Could you outline the two main stages of photosynthesis, including where they take place within the chloroplast, and the primary inputs and outputs for each stage? ... (Human-Bot Dialogue Turns)... (Topic: Photosynthetic Energy Production)
[human]: Please assume the role of an English translator, tasked with correcting and enhancing spelling and language. Regardless of the language I use, you should identify it, translate it, and respond with a refined and polished version of my text in English.
... (Human-Bot Dialogue Turns)... (Topic: Language Translation and Enhancement)
[human]: Suggest five award-winning documentary films with brief background descriptions for aspiring filmmakers to study.
[bot]: ... 5. Än Inconvenient Trutḧ(2006) - Directed by Davis Guggenheim and featuring former United States Vice President Al Gore, this documentary aims to educate the public about global warming. It won two Academy Awards, including Best Documentary Feature. The film is notable for its straightforward yet impactful presentation of scientific data, making complex information accessible and engaging, a valuable lesson for filmmakers looking to tackle environmental or scientific subjects.
... (Human-Bot Dialogue Turns)... (Topic: Documentary Films Recommendation)
[human]: Given the following records of stock prices, extract the highest and lowest closing prices for each month in the year 2022. Return the results as a CSV string, with one line allocated for each month. Date,Open,High,Low,Close,Volume ... ... (Human-Bot Dialogue Turns)... (Topic: Stock Prices Analysis)
[human]: The city of Vega intends to build a bridge that will span the Vegona River, covering a distance of 1.8 kilometers. The proposed location falls within a seismically active area that has experienced several high-magnitude earthquakes. Given these circumstances, what would be the best approach to constructing the bridge? ... (Human-Bot Dialogue Turns)... (Topic: Seismic Bridge Construction)

Question:

Which documentary won two Academy Awards and what lesson does it offer to filmmakers?

Ground-Truth Answer:

"An Inconvenient Truth" (2006) won two Academy Awards. It offers the lesson of making complex information accessible and engaging, a valuable lesson for filmmakers looking to tackle environmental or scientific subjects.

Memory generated by ConditionMem [8]:

- The user is seeking recommendations for award-winning documentary films, along with brief background descriptions, for educational purposes. This request likely stems from a desire to learn from successful examples in the documentary filmmaking field, possibly to improve their own filmmaking skills or to gain inspiration for their projects.
- User is interested in award-winning documentary films for educational purposes, aiming to study their backgrounds and filmmaking techniques.
- The user's request for crafting a succinct and persuasive pitch for a film about overcoming adversity is inspired by the spirit of "March of the Penguins," a documentary that beautifully captures the Emperor penguins' annual journey across the Antarctic, highlighting themes of endurance and survival against the odds. The user seeks to emulate the emotional depth and cinematic quality that made "March of the Penguins" compelling, aiming to create a pitch that resonates with similar themes of resilience and triumph over challenges.
- User is interested in creating content that focuses on themes of overcoming adversity, inspired by the emotional depth and cinematic excellence of documentaries like "March of the Penguins." They value storytelling that captures the essence of endurance and survival against the odds.
... (Other Irrelevant Memory Records) ...

Answer (Conditional Memory):

The documentary that won two Academy Awards is "Free Solo," which follows climber Alex Honnold as he attempts to free solo climb El Capitan in Yosemite National Park. The film offers a lesson to filmmakers on the importance of capturing authentic and compelling human stories, as well as the technical challenges of filming in extreme environments.
(The next part of this figure 9 will be continued on the next page.)

Retrieved Memory (Ours):

[human]: Suggest five award-winning documentary films with brief background descriptions for aspiring filmmakers to study.
[bot]: 1. **"March of the Penguins" (2005)** - Directed by Luc Jacquet, this French feature-length nature documentary delves into the annual journey of Emperor penguins across the Antarctic. It won the Academy Award for Best Documentary Feature. The film is celebrated for its stunning cinematography and its detailed, heartfelt portrayal of the penguins' life cycle, making it a must-study for filmmakers interested in wildlife documentaries. 2. **"Citizenfour" (2014)** ... 5. **"An Inconvenient Truth" (2006)** - Directed by Davis Guggenheim and featuring former United States Vice President Al Gore, this documentary aims to educate the public about global warming. It won two Academy Awards, including Best Documentary Feature. The film is notable for its straightforward yet impactful presentation of scientific data, making complex information accessible and engaging, a valuable lesson for filmmakers looking to tackle environmental or scientific subjects.
[human]: With the spirit in the first film, craft a succinct and persuasive pitch for a film about overcoming adversity.
[bot]: Drawing inspiration from the resilience and majestic journey of Emperor penguins in "March of the Penguins," our documentary, "Unyielding Spirits: The Human Odyssey," will explore the profound stories of individuals around the globe who have faced and overcome staggering adversities. Through breathtaking cinematography and intimate storytelling, we ... (Topic: Documentary Films Recommendation)

Answer (Ours):

"An Inconvenient Truth" won two Academy Awards, including Best Documentary Feature. The film offers a valuable lesson to filmmakers looking to tackle environmental or scientific subjects, as it presents complex information in a straightforward yet impactful way, making it accessible and engaging to a wider audience.

Figure 9: Case study comparing our SeCom with ConditionMem [8]. When using ConditionMem, the agent fails to provide a satisfactory answer due to (1) information loss during the summarization process and (2) the incorrect discarding of turns that are actually useful, as highlighted in brown. The dialogue topic is emphasized in blue. Some context is omitted (replaced by ...) due to length constraints.