MemInsight: Autonomous Memory Augmentation for LLM Agents

Anonymous ACL submission

Abstract

Large language model (LLM) agents have evolved to intelligently process information, make decisions, and interact with users or tools. A key capability is the integration of long-term memory capabilities, enabling these agents to draw upon historical interactions and knowledge. However, the growing memory size and need for semantic structuring pose significant challenges. In this work, we propose an autonomous memory augmentation approach, MemInsight, to enhance semantic data representation and retrieval mechanisms. By leveraging autonomous augmentation to historical interactions, LLM agents are shown to deliver more accurate and contextualized responses. We empirically validate the efficacy of our proposed approach in three task scenarios; conversational recommendation, question answering and event summarization. On the LLM-REDIAL dataset, MemInsight boosts persuasiveness of recommendations by up to 14%. Moreover, it outperforms a RAG baseline by 34% in recall for LoCoMo retrieval. Our empirical results show the potential of MemInsight to enhance the contextual performance of LLM agents across multiple tasks[1].

1 Introduction

LLM agents have emerged as an advanced framework to extend the capabilities of LLMs to improve reasoning (Yao et al., 2023; Wang et al., 2024c), adaptability (Wang et al., 2024d), and self-evolution (Zhao et al., 2024a; Wang et al., 2024e; Tang et al., 2025). A key component of these agents is their memory module, which retains past interactions to allow more coherent, consistent, and personalized responses across various tasks. The memory of the LLM agent is designed to emulate human cognitive processes by simulating how knowledge is accumulated and historical experiences are leveraged to facilitate complex reasoning and the retrieval of relevant information to inform actions (Zhang et al., 2024). However, the advantages of an LLM agent's memory also introduce notable challenges (Wang et al., 2024b). (1) As data accumulates over time, retrieving relevant information becomes increasingly challenging, especially during extended interactions or complex tasks. (2) Processing large historical data, which can grow rapidly, as interactions accumulate, requires effective memory management strategies. (3) Storing data in its raw format can hinder efficient retrieval of pertinent knowledge, as distinguishing between relevant and irrelevant details becomes more challenging, potentially leading to noisy or imprecise information that compromises the agent's performance. Furthermore, (4) the integration of knowledge across tasks is constrained, limiting the agent's ability to effectively utilize data from diverse contexts. Consequently, effective knowledge representation and structuring of LLM agent memory are essential to accumulate relevant information and enhance understanding of past events. Improved memory management enables better retrieval and contextual awareness, making this a critical and evolving area of research.

Hence, in this paper we introduce an autonomous memory augmentation approach, MemInsight, which empowers LLM agents to identify critical information within the data and proactively propose effective attributes for memory enhancements. This is analogous to the human processes of attentional control and cognitive updating, which involve selectively prioritizing relevant information, filtering out distractions, and continuously refreshing the mental workspace with new and pertinent data (Hu et al., 2024; Hou et al., 2024).

MemInsight autonomously generates augmentations that encode both relevant semantic and contextual information for memory. These aug-

[1]The source code and data samples will be released after paper acceptance.https://github.com/anonymous.

mentations facilitate the identification of memory components pertinent to various tasks. Accordingly, MemInsight can improve memory retrieval by leveraging relevant attributes of memory, thereby supporting autonomous LLM agent adaptability and self-evolution.

Our contributions can be summarized as follows:

- We propose a structured autonomous approach that adapts LLM agents' memory representations while preserving context across extended conversations for various tasks.
- We design and apply memory retrieval methods that leverage the generated memory augmentations to filter out irrelevant memory while retaining key historical insights.
- Our promising empirical findings demonstrate the effectiveness of MemInsight on several tasks: conversational recommendation, question answering, and event summarization.

2 Related Work

Well-organized and semantically rich memory structures enable efficient storage and retrieval of information, allowing LLM agents to maintain contextual coherence and provide relevant responses. Developing an effective memory module in LLM agents typically involves two critical components: structural memory generation and memory retrieval methods (Zhang et al., 2024; Wang et al., 2024a).

LLM Agents Memory Recent research in LLM agents memory focuses on developing methods for effectively storing previous interactions and feedback (Packer et al., 2024). Contemporary approaches emphasize memory structures that enhance the adaptability of agents and improve their ability to generalize to previously unseen environments (Zhao et al., 2024a; Zhang et al., 2024; Zhu et al., 2023). Common memory forms include summaries and abstract high-level information from raw observations to capture key points and reduce information redundancy (Maharana et al., 2024). Other approaches include structuring memory as summaries, temporal events, or reasoning chains (Zhao et al., 2024a; Zhang et al., 2024; Zhu et al., 2023; Maharana et al., 2024; Anokhin et al., 2024; Liu et al., 2023a). In addition, there are studies that enrich raw conversations with semantic representations like sequence of events and historical event summaries (Zhong et al., 2023; Maharana et al., 2024) or extract reusable workflows from canonical examples and integrate them into memory to assist test-time inference (Wang et al., 2024f). However, all aforementioned studies rely on either unstructured memory or human-designed attributes for memory representation, while MemInsight leverages the AI agent's autonomy to discover the ideal attributes for structured representation.

LLM Agents Memory Retrieval Existing works have leveraged memory retrieval techniques for efficiency when tackling vast amounts of historical context (Hu et al., 2023a; Zhao et al., 2024b; Tack et al., 2024; Ge et al., 2025). Common approaches for memory retrieval include generative retrieval models, which encode memory as dense vectors and retrieve the top-k relevant documents based on similarity search techniques (Zhong et al., 2023; Penha et al., 2024). Various similarity metrics, such as cosine similarity (Packer et al., 2024), are employed, alongside advanced techniques like dual-tower dense retrieval models, which encode each memory history into embeddings indexed by FAISS (Johnson et al., 2017) to enhance retrieval efficiency (Zhong et al., 2023). Additionally, methods such as Locality-Sensitive Hashing (LSH) are utilized to retrieve tuples containing related entries in memory (Hu et al., 2023b).

3 Autonomous Memory Augmentation

Our proposed MemInsight model is designed to enhance memory representation through a structured augmentation process that optimizes memory retrieval. Figure 1 presents an overview of the model, highlighting its main modules: attribute mining, annotation, and memory retriever.

3.1 Attribute Mining and Annotation

To ensure the effectiveness of these attributes in future interactions, they must be meaningful, accurate, and Attribute mining in our MemInsight model leverages a backbone LLM to autonomously identify and define key attributes that encapsulate semantic knowledge from user interactions. This entails selecting attributes most relevant to the task under consideration and employs them to annotate historical conversations. Effective attributes must be meaningful, accurate, and contextually relevant to enhance future interactions. To achieve this, the augmentation process follows a structured approach, defining the perspective from which the

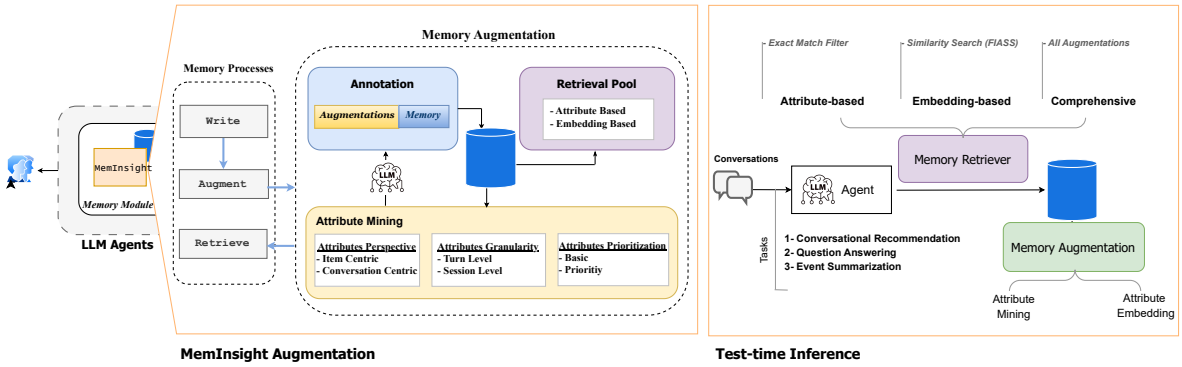

Figure 1: Main modules of MemInsight including, Attribute Mining, Memory Retrieval, and Annotation, triggered by different memory processes: Augment, Write, and Retrieve. In addition to the test-time inference evaluation downstream tasks, memory augmentation and adopted memory retrieval methods.

attributes are derived, determining the appropriate level of augmentation granularity, and establishing a coherent sequence for annotation. The resulting attributes and values are then used to enrich memory, ensuring a well-organized and informative representation of past interactions.

3.1.1 Attribute Perspective

Attribute generation is guided by two primary orientations: entity-centric and conversation-centric. Entity-centric emphasizes a specific item stored in memory such as movies or books. Attributes generated for entity-centric augmentations should capture the main characteristics and features of this entity. For example, attributes for a movie entity might include the director, actors, and year of release, while attributes for a book entity would encompass the author, publisher, and number of pages. On the other hand, conversation-centric augmentations focus on annotating and characterizing the entire user interaction from the user's perspective. This approach ensures that the extracted attributes align with the user's intent, preferences, sentiment, emotions, motivations, and choices, thereby improving personalized responses and memory retrieval. An illustrative example is provided in Figure 4.

3.1.2 Attribute Granularity

While entity-centric augmentations focus on specific entities in memory, conversation-centric augmentations introduce an additional factor: attribute granularity, which determines the level of details captured in the augmentation process. The augmentation attributes can be analyzed at varying levels of abstraction, either at the level of individual turns within a user conversation (turn-level), or across the entire dialogue session (session-level), each offering distinct insights into the conversational context. At the turn level, each dialogue turn is independently augmented, focusing on the specific content of individual turns to generate more nuanced and contextual attributes. In contrast, session-level annotation considers the entire dialogue, generating generalized attributes that capture the broader conversational context. Due to its broader granularity, session-level augmentation emphasizes high-level attributes and conversational structures rather than the detailed features of individual turns. An example of both levels is illustrated in Figure 2 for a sample dialogue turns. As shown, turn-level annotations offer finer-grained details, while session-level annotations provide a broader overview of the dialogue.

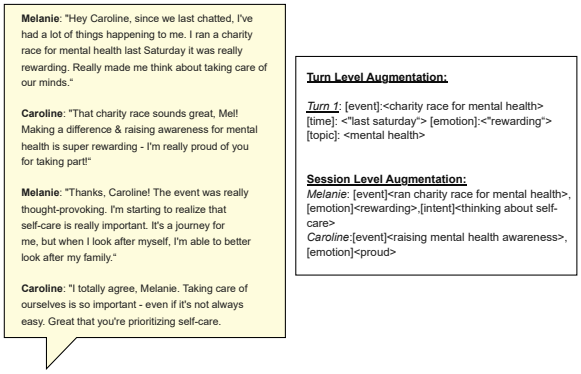

Figure 2: An example for Turn level and Session level annotations for a sample dialogue conversation from the LoCoMo Dataset.

3.1.3 Annotation and Attribute Prioritization

Subsequently, the generated attributes and their corresponding values are used to annotate the agent's memory. Annotation is done by aggregating attributes and values in the relevant memory in the

form:

$$\{m_i : \langle a_1, v_1 \rangle, ..., \langle a_n, v_n \rangle\} \quad (1)$$

where m_i stands for relevant memory and a_i, v_i denote attributes and values respectively. The relevant memory may correspond to turn or session level. Attributes are typically aggregated using the Attribute Prioritization method, which can be classified into Basic and Priority. In Basic Augmentation, attributes are aggregated without a predefined order, resulting in an arbitrary sequence $i_1, .., i_n$. In contrast, Priority Augmentation sorts attribute-value pairs according to their relevance to the memory being augmented. This prioritization follows a structured order in which attribute i_1 holds the highest significance, ensuring that more relevant attributes are processed first.

3.2 Memory Retrieval

MemInsight augmentations are employed to enrich or retrieve relevant memory. For comprehensive retrieval, memory is retrieved along with all associated augmentations to generate a more context-aware response. Additionally, MemInsight can refine the retrieval process. Initially, the current context is augmented to identify task-specific and interaction-related attributes, which then guide the retrieval of the pertinent memory. Two primary retrieval methods are proposed: (1) *Attribute-based Retrieval*, leverages the current context to generate attributes tailored to the specific task at hand. These attributes serve as criteria for selecting and retrieving relevant memory that shares similar attributes in their augmentations. The retrieved memories, which align with the required attributes, are subsequently integrated into the current context to enrich the ongoing interaction. (2) *Embedding-based Retrieval*, utilizes memory augmentations to create a unique embedding representation for each memory instance, derived from its aggregated annotations. Simultaneously, the augmentations of the current context are embedded to form a query vector, which is then used in a similarity-based search to retrieve the top-k most relevant memories. Finally, all retrieved memory are incorporated into the current context to enhance the relevance and coherence of the ongoing interaction. A detailed description of this method can be found in Appendix C.

4 Evaluation

4.1 Datasets

We conduct a series of experiments on the datasets: LLM-REDIAL (Liang et al., 2024) and LoCoMo (Maharana et al., 2024). LLM-REDIAL is a dataset for evaluating movie Conversational Recommendation, containing approximately 10K dialogues covering 11K movies in memory. While LoCoMo is a dataset for evaluating Question Answering and Event Summarization, consisting of 30 long-term dialogues across up to 10 sessions between two speakers. LoCoMo includes five question categories: Single-hop, Multi-hop, Temporal reasoning, Open-domain knowledge, and Adversarial questions. Each question has a reference label that specifies the relevant dialogue turn in memory required to generate the answer. Additionally, LoCoMo provides event labels for each speaker in a session, which we use as ground truth for Event Summarization evaluation.

4.2 Experimental Setup

To evaluate our model, we begin by augmenting the datasets using a backbone LLM with zero-shot prompting to identify relevant attributes and their corresponding values. For augmentation generation and evaluation across various tasks, we utilize the following models for attribute generation: Claude Sonnet,[2] Llama[3] and Mistral.[4] For the Event Summarization task, we also use the Claude-3-Haiku model.[5] For embedding-based retrieval tasks, we employ the Titan Text Embedding model [6] to generate embeddings. The augmented memory is then embedded and indexed using FAISS (Johnson et al., 2017) for vector indexing and search. To ensure consistency across all experiments, we use the same base model for the primary tasks: recommendation, answer generation, and summarization, while evaluating different models for augmentation. Claude Sonnet serves as the backbone LLM in baselines for all tasks.

4.3 Evaluation Metrics

The evaluation metrics used for assessing different tasks using MemInsight include, traditional metrics like F1-score metric for answer prediction and

[2] claude-3-sonnet-20240229-v1
[3] llama3-70b-instruct-v1
[4] mistral-7b-instruct-v0
[5] claude-3-Haiku-20240307-v1
[6] titan-embed-text-v2:0

recall for accuracy in Question Answering. Recall@K and NDCG@K for Conversational Recommendation, along with LLM-based metrics for genre matching.

We also evaluate using subjective metrics including Persuasiveness, used in Liang et al. (2024), to assess how persuasive the recommendations are relative to the ground truth. Additionally, we introduce a Relatedness metric, where we prompt an LLM to measure how comparable the recommendation attributes are to the ground truth, categorizing them as not comparable, comparable, or highly comparable. Finally, we assess Event Summarization using an LLM-based metric, G-Eval (Liu et al., 2023b), a summarization evaluation metric that measures the relevance, consistency, and coherence of generated summaries as opposed to reference labels. These metrics provide a comprehensive framework for evaluating both retrieval effectiveness and response quality.

5 Experiments

5.1 Questioning Answering

Questioning Answering task experiments are conducted to evaluate the effectiveness of MemInsight in answer generation. We assess overall accuracy to measure the system's ability to retrieve and incorporate relevant information from augmentations. The base model, which incorporates all historical dialogues without augmentation, serves as the baseline. We additionally consider Dense Passage Retrieval (DPR) RAG model (Karpukhin et al., 2020) as a comparative baseline due to its speed and scalability.

Memory Augmentation In this task, memory is constructed from historical conversational dialogues, which requires the generation of conversation-centric attributes for augmentation. Given that the ground truth labels consist of dialogue turns relevant to the question, the dialogues are annotated at the turn level. A backbone LLM is prompted to generate augmentation attributes for both conversation-centric and turn-level annotations.

Memory Retrieval To answer a given question, the relevant dialogue turn must be retrieved from historical dialogues. In order to retrieve the relevant dialogue turn, the question is first augmented to identify relevant attributes and a memory retrieval method is applied. We evaluate different MemInsight memory retrieval methods to demonstrate the efficacy of our model. We employ attribute-based retrieval by selecting dialogue turns augmented with attributes that exactly match the question's attributes. Additionally, we evaluate the embedding-based retrieval, where the augmentations are embedded and indexed for retrieval. Hence, the question and attributes are transformed into an embedded query, which is used to perform a vector similarity search to retrieve the top-k most similar dialogue turns. Once the relevant memory is retrieved, it is integrated into the current context to generate the final answer.

Experimental Results We initiate our evaluation by assessing attribute-based memory retrieval using the Claude-3-Sonnet model. Table 1 presents the overall F1 score, measuring the accuracy of the generated answers. As shown in the table, attribute-based retrieval outperforms the baseline model by 3% in overall accuracy, with notable improvements in single-hop, temporal reasoning, and adversarial questions, which require advanced contextual understanding and reasoning. These results indicate that the augmented history enriched the context, leading to better reasoning and a significant increase in the F1 score for answer generation. Additionally, we perform a detailed analysis of embedding-based retrieval, where we consider evaluating basic and priority augmentation using the Claude-3-Sonnet model.

Table 1 demonstrates that the priority augmentation consistently outperforms the basic model across all questions. This finding suggests that the priority relevance of augmentations enhances context representation for conversational data. Subsequently, we evaluate the priority augmentations using Llama, and Mistral models for Embedding-based retrieval. As shown in the table, the Embedding-based retrieval outperforms the RAG baseline across all question categories, except for adversarial questions, yet the overall accuracy of MemInsight remains superior. Additionally, MemInsight demonstrates a significant improvement in performance on multi-hop questions, which require reasoning over multiple pieces of supporting evidence. This suggests that the generated augmentations provided a more robust understanding and a broader perspective of the historical dialogues. RECALL metrics in Table 2 revealed a more significant boost, with priority augmentations increasing accuracy across all categories and

Model	Single-hop	Multi-hop	Temporal	Open-domain	Adversarial	Overall
Baseline (Claud-3-Sonnet)	15.0	10.0	3.3	26.0	45.3	26.1
Attribute-based Retrieval						
MemInsight (Claude-3-Sonnet)	**18.0**	**10.3**	**7.5**	**27.0**	**58.3**	**29.1**
Embedding-Based Retrieval						
RAG Baseline (DPR)	11.9	9.0	6.3	12.0	**89.9**	28.7
MemInsight (Llama v3$_{Priority}$)	14.3	13.4	6.0	15.8	82.7	29.7
MemInsight (Mistral v1$_{Priority}$)	**16.1**	14.1	6.1	16.7	81.2	30.0
MemInsight (Claude-3-Sonnet$_{Basic}$)	14.7	13.8	5.8	15.6	82.1	29.6
MemInsight (Claude-3-Sonnet$_{Priority}$)	15.8	**15.8**	**6.7**	**19.1**	75.3	**30.1**

Table 1: Results for F1 Score (%) for answer generation accuracy for attribute-based and embedding-based memory retrieval methods. Baseline is Claude-3-Sonnet model to generate answers using all memory without augmentation, for Attribute-based retrieval. In addition to the Dense Passage Retrieval(DPR) for Embedding-based retrieval. Evaluation is done with $k = 5$. Best results per question category over all methods are in bold.

Model	Single-hop	Multi-hop	Temporal	Open-domain	Adversarial	Overall
RAG Baseline (DPR)	15.7	31.4	15.4	15.4	34.9	26.5
MemInsight (Llama v3$_{Priority}$)	31.3	63.6	23.8	53.4	28.7	44.9
MemInsight (Mistral v1$_{Priority}$)	31.4	63.9	26.9	58.1	36.7	48.9
MemInsight (Claude-3-Sonnet$_{Basic}$)	33.2	67.1	29.5	56.2	35.7	48.8
MemInsight (Claude-3-Sonnet$_{Priority}$)	**39.7**	**75.1**	**32.6**	**70.9**	**49.7**	**60.5**

Table 2: Results for the RECALL@k=5 accuracy for Embedding-based retrieval for answer generation using LoCoMo dataset. Dense Passage Retrieval(DPR) RAG model is the baseline. Best results are in bold.

Statistic		Count
Total Movies		9687
Avg. Attributes		7.39
Failed Attributes		0.10%
Top-5 Attributes	Genre	9662
	Release year	5998
	Director	5917
	Setting	4302
	Characters	3603

Table 3: Statistics of attributes generated for the LLM-REDIAL Movie dataset, which include total number of movies, average number of attributes per item, number of failed attributes, and the counts for the most frequent five attributes.

yielding a 35% overall improvement.

5.2 Conversational Recommendation

We simulate conversational recommendation by preparing dialogues for evaluation under the same conditions proposed by Liang et al. (2024). This process involves masking the dialogue and randomly selecting $n = 200$ conversations for evaluation to ensure a fair comparison. Each conversational dialogue used is processed by masking the ground truth labels, followed by a turn cut-off, where all dialogue turns following the first masked turn are removed and retained as evaluation labels. Subsequently, the dialogues are augmented using a conversation-centric approach to identify relevant user interest attributes for retrieval. Finally, we prompt the LLM model to generate a movie recommendation that best aligns with the masked token, guided by the augmented movies retrieved based on the user's historical interactions.

The baseline for this evaluation is the results presented in the LLM-REDIAL paper (Liang et al., 2024) which employs zero-shot prompting for recommendation using the ChatGPT model[7]. In addition to the baseline model that uses memory without augmentation.

Evaluation includes direct matches between recommended and ground truth movie titles using RECALL@[1,5,10] and NDCG@[1,5,10]. Furthermore, to address inconsistencies in movie titles generated by LLMs, we incorporate an LLM-based evaluation that assesses recommendations based on genre similarity. Specifically, a recommended movie is considered a valid match if it shares the same genre as the corresponding ground truth label.

Memory Augmentation We initially augment the dataset with relevant attributes, primarily employing entity-centric augmentations for memory annotation, as the memory consists of movies. In this context, we conduct a detailed evaluation of the generated attributes to provide an initial assessment of the effectiveness and relevance of MemInsight augmentations. To evaluate the quality of the generated attributes, Table 3 presents statistical data on the generated attributes, including the five most frequently occurring attributes across the entire dataset. As shown in the table, the generated attributes are generally relevant, with "genre" being the most significant attribute based on its cumulative frequency across all movies (also shown in Figure 5). However, the relevance of attributes vary, emphasizing the need for prioritization in augmentation. Additionally, the table reveals that

[7]https://openai.com/blog/chatgpt

Model	Avg. Items Retrieved	Direct Match (↑)			Genre Match (↑)			NDCG(↑)		
		R@1	R@5	R@10	R@1	R@5	R@10	N@1	N@5	N@10
Baseline (Claude-3-Sonnet)	144	0.000	0.010	0.015	0.320	0.57	0.660	0.005	0.007	0.008
LLM-REDIAL Model	144	-	0.000	0.005	-	-	-	-	0.000	0.001
Attribute-Based Retrieval										
MemInsight (Claude-3-Sonnet)	15	0.005	0.015	0.015	0.270	0.540	0.640	0.005	0.007	0.007
Embedding-Based Retrieval										
MemInsight (Llama v3)	10	0.000	0.005	**0.028**	0.380	0.580	0.670	0.000	0.002	0.001
MemInsight (Mistral v1)	10	0.005	0.010	0.010	0.380	0.550	0.630	0.005	0.007	0.007
MemInsight (Claude-3-Haiku)	10	0.005	0.010	0.010	0.360	**0.610**	0.650	0.005	0.007	0.007
MemInsight (Claude-3-Sonnet)	10	0.005	0.015	0.015	**0.400**	0.600	0.64	0.005	0.010	0.010
Comprehensive										
MemInsight (Claude-3-Sonnet)	144	**0.010**	**0.020**	0.025	0.300	0.590	**0.690**	**0.010**	**0.015**	**0.017**

Table 4: Results for Movie Conversational Recommendation using (1) Attribute-based retrieval with Claude-3-Sonnet model (2) Embedding-based retrieval across models (Llama v3, Mistral v1, Claude-3-Haiku, and Claude-3-Sonnet) (3) Comprehensive setting using Claude-3-Sonnet that includes **ALL** augmentations. Evaluation metrics include RECALL, NDCG, and an LLM-based genre matching metric, with $n = 200$ and $k = 10$. Baseline is Claude-3-Sonnet without augmentation. Best results are in bold.

augmentation was unsuccessful for 0.1% of the movies, primarily due to the LLM's inability to recognize certain movie titles or because the presence of some words in the movie titles conflicted with the LLM's policy.

Memory Retrieval For this task we evaluate attribute-based retrieval using the Claude-3-Sonnet model with both filtered and comprehensive settings. Additionally, we examine embedding-based retrieval using all other models. For embedding-based retrieval, we set $k = 10$, meaning that 10 memory instances are retrieved (as opposed to 144 in the baseline).

Experimental Results Table 4 shows the results for conversational recommendation evaluating comprehensive setting, attribute-based retrieval and embedding-based retrieval. As shown in the table, comprehensive memory augmentation tends to outperform the baseline and LLM-REDIAL model for recall and NDCG metrics. For genre match we find the results to be comparable when considering all attributes. However, attributed-based filtering retrieval still outperforms the LLM-REDIAL model and is comparable to the baseline with almost 90% less memory retrieved.

Table 5 presents the results of subjective LLM-based evaluation for Persuasiveness and Relatedness. The findings indicate that memory augmentation enhances partial persuasiveness by 10–11% using both comprehensive and attribute-based retrieval, while also reducing unpersuasive recommendations and increasing highly persuasive ones by 4% in attribute-based retrieval. Furthermore, the results highlights the effectiveness of embedding-based retrieval, which leads to a 12% increase in highly persuasive recommendations and enhances all relatedness metrics. This illustrates how MemInsight enriches the recommendation process by incorporating condensed, relevant knowledge, thereby producing more persuasive and related recommendations. However, these improvements were not reflected in recall and NDCG metrics.

5.3 Event Summarization

We evaluate the effectiveness of MemInsight in enriching raw dialogues with relevant insights for event summarization. We utilize the generated annotations to identify key events within conversations and hence use them for event summarization. We compare the generated summaries against LoCoMo's event labels as the baseline. Figure 3 illustrates the experimental framework, where the baseline is the raw dialogues sent to the LLM model to generate an event summary, then both event summaries, from raw dialogues and augmentation based summaries, are compared to the ground truth summaries in the LoCoMo dataset.

Memory Augmentation In this experiment, we evaluate the effectiveness of augmentation granularity; turn-level dialogue augmentations as opposed to session-level dialogue annotations. We additionally, consider studying the effectiveness of using only the augmentations to generate the event summaries as opposed to using both the augmentations and their corresponding dialogue content.

Experimental Results As shown in Table 6, our MemInsight model achieves performance comparable to the baseline, despite relying only on dialogue turns or sessions containing the event label. Notably, turn-level augmentations provided more

Model	Avg. Items Retrieved	LLM-Persuasiveness %			LLM-Relatedness%		
		Unpers*	Partially Pers.	Highly Pers.	Not Comp*	Comp	Match
Baseline (Claude-3-Sonnet)	144	16.0	64.0	13.0	57.0	41.0	2.0
Attribute-Based Retrieval							
MemInsight (Claude-3-Sonnet)	15	2.0	**75.0**	17.0	40.5	54.0	2.0
Embedding-Based Retrieval							
MemInsight (Llama v3)	10	11.3	63.0	20.4	19.3	80.1	0.5
MemInsight (Mistral v1)	10	16.3	61.2	18.0	**16.3**	**82.5**	**5.0**
MemInsight (Claude-3-Haiku)	10	**1.6**	53.0	**25.0**	23.3	74.4	2.2
MemInsight (Claude-3-Sonnet)	10	2.0	59.5	20.0	29.5	68.0	2.5
Comprehensive							
MemInsight (Claude-3-Sonnet)	144	2.0	74.0	12.0	42.5	56.0	1.0

Table 5: Movie Recommendations results (with similar settings to Table 4) using LLM-based metrics; (1) Persuasiveness— % of Unpersuasive (lower is better), Partially, and Highly Persuasive cases. (2) Relatedness— % of Not Comparable (lower is better), Comparable, and Exactly Matching cases. Best results are in bold. Comprehensive setting includes **ALL** augmentations. Totals may NOT sum to 100% due to cases the LLM model could not evaluate.

Model	Claude-3-Sonnet			Llama v3			Mistral v1			Claude-3-Haiku		
	Rel.	Coh.	Con.	Rel.	Coh.	Con.	Rel.	Coh.	Con.	Rel.	Coh.	Con.
Baseline Summary	3.27	**3.52**	2.86	2.03	2.64	2.68	3.39	3.71	4.10	4.00	4.4	3.83
MemInsight (TL)	3.08	3.33	2.76	1.57	2.17	1.95	2.54	2.53	2.49	3.93	4.3	3.59
MemInsight (SL)	3.08	3.39	2.68	2.0	2.62	3.67	4.13	4.41	4.29	3.96	4.30	3.77
MemInsight +Dialogues (TL)	**3.29**	3.46	**2.92**	**2.45**	2.19	2.87	**4.30**	**4.53**	**4.60**	**4.23**	**4.52**	**4.16**
MemInsight +Dialogues (SL)	3.05	3.41	2.69	2.24	**2.80**	**3.86**	4.04	4.48	4.33	3.93	4.33	3.73

Table 6: Event Summarization results using G-Eval metrics (higher is better): Relevance, Coherence, and Consistency. Comparing summaries generated with augmentations only at Turn-Level (TL) and Session-Level (SL) and summaries generated using both augmentations and dialogues (MemInsight +Dialogues) at TL and SL. Best results are in bold.

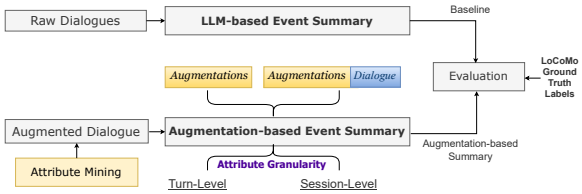

Figure 3: Evaluation framework for event summarization with MemInsight, exploring augmentation at Turn and Session levels, considering attributes alone or both attributes and dialogues for richer summaries.

Model	G-Eval % (↑)		
	Rel.	Coh.	Con.
Baseline(Llama v3)	2.03	2.64	2.68
Llama v3 + Llama v3	2.45	2.19	2.87
Claude-3-Sonnet + Llama v3	**3.15**	**3.59**	**3.17**

Table 7: Results for Event Summarization using Llama v3, where the baseline is the model without augmentation as opposed to the augmentation model (turn-level) using Claude-3-Sonnet vs Llama v3.

precise and detailed event information, leading to improved performance over both the baseline and session-level annotations.

For Claude-3-Sonnet, all metrics remain comparable, indicating that memory augmentations effectively capture the semantics and knowledge within dialogues at both the turn and session levels. This proves that the augmentations sufficiently enhance context representation for generating event summaries.

To further investigate how backbone LLMs impact augmentation quality, we employed Claude-3-Sonnet as opposed to Llama v3 for augmentation while still using Llama for event summarization. As presented in Table 7, Sonnet augmentations resulted in improved performance for all metrics, providing empirical evidence for the effectiveness and stability of Sonnet in augmentation.

6 Conclusion

This paper introduced MemInsight, an autonomous memory augmentation method that enhances LLM agents memory through attribute-based annotations. While maintaining comparable performance on standard metrics, MemInsight significantly improves LLM-based evaluation scores, highlighting its effectiveness in capturing semantics and boosting performance across tasks and datasets. Additionally, attribute-based filtering and embedding retrieval methods showed promising methods of utilizing the generated augmentations to improve the performance of various tasks. Priority augmentation enhancing similarity searches and retrieval. MemInsight also could be a complement to RAG models for customized retrievals, integrating LLM knowledge. Results confirm that attribute-based retrieval effectively enriches recommendation tasks, leading to more persuasive recommendations.

7 Limitations

While the proposed MemInsight model demonstrates significant potential in enhancing retrieval and contextual understanding, certain limitations must be acknowledged. MemInsight relies on the quality and granularity of annotations generated using LLMs, making it susceptible to issues such as hallucinations inherent to LLM outputs. Furthermore, although the current evaluation metrics provide valuable insights, they may not comprehensively capture all aspects of retrieval and generation quality, highlighting the need for the development of more robust and multidimensional evaluation frameworks.

References

Petr Anokhin, Nikita Semenov, Artyom Sorokin, Dmitry Evseev, Mikhail Burtsev, and Evgeny Burnaev. 2024. Arigraph: Learning knowledge graph world models with episodic memory for llm agents. *Preprint*, arXiv:2407.04363.

Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu, Monica Sunkara, Yassine Benajiba, and Yi Zhang. 2025. Tremu: Towards neuro-symbolic temporal reasoning for llm-agents with memory in multi-session dialogues. *arXiv preprint arXiv:2502.01630*.

Yuki Hou, Haruki Tamoto, and Homei Miyashita. 2024. "my agent understands me better": Integrating dynamic human-like memory recall and consolidation in llm-based agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, page 1–7. ACM.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023a. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*.

Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. 2023b. Chatdb: Augmenting llms with databases as their symbolic memory. *Preprint*, arXiv:2306.03901.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2024. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. *Preprint*, arXiv:2408.09559.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *Preprint*, arXiv:1702.08734.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. LLM-REDIAL: A large-scale dataset for conversational recommender systems created from user behaviors with LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8926–8939, Bangkok, Thailand. Association for Computational Linguistics.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 2023a. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *Preprint*, arXiv:2311.08719.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Memgpt: Towards llms as operating systems. *Preprint*, arXiv:2310.08560.

Gustavo Penha, Ali Vardasbi, Enrico Palumbo, Marco de Nadai, and Hugues Bouchard. 2024. Bridging search and recommendation in generative retrieval: Does one task help the other? *Preprint*, arXiv:2410.16823.

Jihoon Tack, Jaehyung Kim, Eric Mitchell, Jinwoo Shin, Yee Whye Teh, and Jonathan Richard Schwarz. 2024. Online adaptation of language models with a memory of amortized contexts. *arXiv preprint arXiv:2403.04317*.

Zhengyang Tang, Ziniu Li, Zhenyang Xiao, Tian Ding, Ruoyu Sun, Benyou Wang, Dayiheng Liu, Fei Huang, Tianyu Liu, Bowen Yu, and Junyang Lin. 2025. Enabling scalable oversight via self-evolving critic. *Preprint*, arXiv:2501.05727.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024a. Mixture-of-agents enhances large language model capabilities. *Preprint*, arXiv:2406.04692.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024b. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6).

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024c. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *Preprint*, arXiv:2402.18272.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024d. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *Preprint*, arXiv:2402.18272.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024e. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *Preprint*, arXiv:2402.18272.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2024f. Agent workflow memory. *Preprint*, arXiv:2409.07429.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *Preprint*, arXiv:2404.13501.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024a. Expel: Llm agents are experiential learners. *Preprint*, arXiv:2308.10144.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024b. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19632–19642.

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

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *Preprint*, arXiv:2305.17144.

A Ethical Consideration

We have thoroughly reviewed the licenses of all scientific artifacts, including datasets and models, ensuring they permit usage for research and publication purposes. To protect anonymity, all datasets used are de-identified. Our proposed method demonstrates considerable potential in significantly reducing both the financial and environmental costs typically associated with enhancing large language models. By lessening the need for extensive data collection and human labeling, our approach not only streamlines the process but also provides an effective safeguard for user and data privacy, reducing the risk of information leakage during training corpus construction. Additionally, throughout the paper-writing process, Generative AI was exclusively utilized for language checking, paraphrasing, and refinement.

B Autonomous Memory Augmentation

B.1 Attribute Mining

Figure 4 illustrates examples for the two types of attribute augmentation: entity-centric and conversation-centric. The entity-centric augmentation represents the main attributes generated for the book entitled 'Already Taken', where attributes are derived based on entity-specific characteristics such as genre, author, and thematic elements. The conversation-centric example illustrates the augmentation generated for a sample two turns dialogue from the LLM-REDIAL dataset, highlighting attributes that capture contextual elements such as user intent, motivation, emotion, perception, and genre of interest.

Furthermore, Figure 5 presents an overview of the top five attributes across different domains in the LLM-REDIAL dataset. These attributes represent the predominant attributes specific to each domain, highlighting the significance of different attributes in augmentation generation. Consequently, the integration of priority-based embeddings has led to improved performance.

C Embedding-based Retrieval

In the context of embedding-based memory retrieval, movies are augmented using MemInsight, and the generated attributes are embedded to retrieve relevant movies from memory. Two main embedding methods were considered:

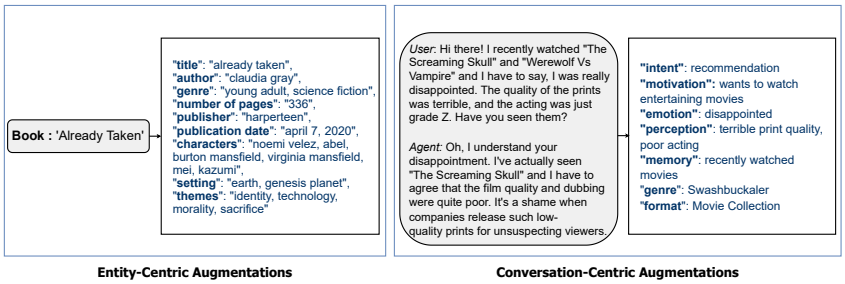

Figure 4: An example of entity-centric augmentation for the book 'Already Taken', and a conversation-centric augmentation for a sample dialogue from the LLM-REDIAL dataset.

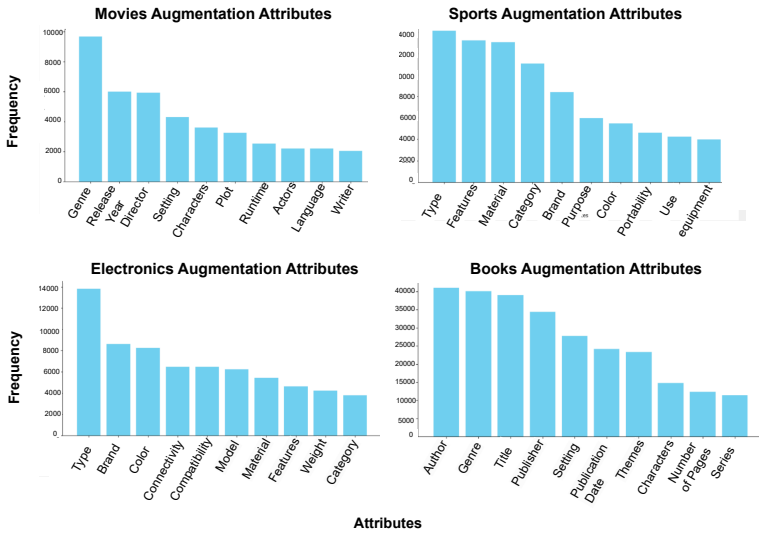

Figure 5: Top 10 attributes by frequency in the LLM-REDIAL dataset across domains (Movies, Sports Items, Electronics, and Books) using MemInsight Attribute Mining. Frequency indicates how often each attribute was generated to augment different movies.

(1) Averaging Over Independent Embeddings Each attribute and its corresponding value in the generated augmentations is embedded independently. The resulting attribute embeddings are then averaged across all attributes to generate the final embedding vector representation, as illustrated in Figure 6 which are subsequently used in similarity search to retrieve relevant movies.

(2) All Augmentations Embedding In this method, all generated augmentations, including all attributes and their corresponding values, are encoded into a single embedding vector and stored for retrieval as shown in Figure 6. Additionally, Figure 7 presents the cosine similarity results for both methods. As depicted in the figure, averaging over all augmentations produces a more consistent and reliable measure, as it comprehensively captures all attributes and effectively differentiates between similar and distinct characteristics. Consequently, this method was adopted in our experiments.

D Question Answering

D.1 Prompts

Table 8 outlines the prompts used in the Question Answering task for generating augmentations in both questions and conversations.

E Conversational Recommendation

E.1 Prompts

Table 9 presents the prompts used in Conversational Recommendation for movie recommendations, incorporating both basic and priority augmentations.

E.2 Evaluation Framework

Figure 8 presents the evaluation framework for the Conversation Recommendation task. The process begins with (1) augmenting all movies in memory

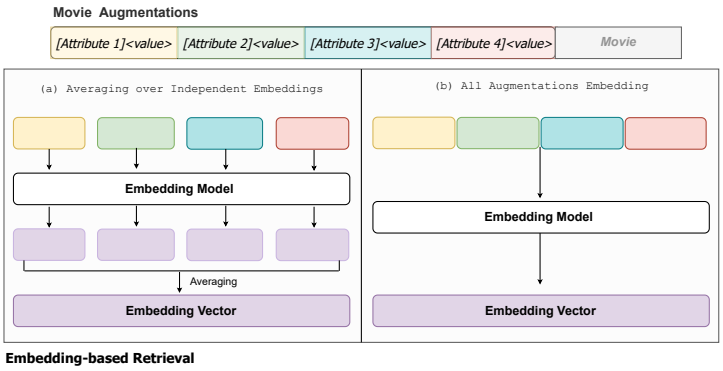

Figure 6: Embedding methods for Embedding-based retrieval methods using generated Movie augmentations including (a) Averaging over Independent Embeddings and (b) All Augmentations Embedding.

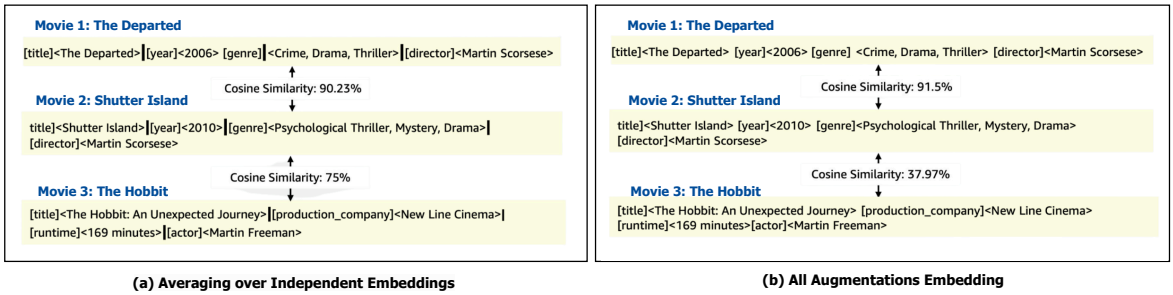

Figure 7: An illustrative example of augmentation embedding methods for three movies: (1) The Departed, (2) Shutter Island, and (3) The Hobbit. Movies 1 and 2 share similar attributes, whereas movies 1 and 3 differ. Te top 5 attributes of every movie were selected for a simplified illustration.

using entity-centric augmentations to enhance retrieval effectiveness. (2) Next, all dialogues in the dataset are prepared to simulate the recommendation process by masking the ground truth labels and prompting the LLM to find the masked labels based on augmentations from previous user interactions. (3) Recommendations are then generated using the retrieved memory, which may be attribute-based—for instance, filtering movies by specific attributes such as genre or using embedding-based retrieval. (4) Finally, the recommended movies are evaluated against the ground truth labels to assess the accuracy and effectiveness of the retrieval and recommendation approach.

E.3 Event Summarization

E.3.1 Prompts

Table 10 presents the prompt used in Event Summarization to augment dialogues by generating relevant attributes. In this process, only attributes related to events are considered to effectively summarize key events from dialogues, ensuring a focused and structured summarization approach.

F Qualitative Analysis

Figure 9 illustrates the augmentations generated using different LLM models, including Claude-Sonnet, Llama, and Mistral for a dialogue turn from the LoCoMo dataset. As depicted in the figure, augmentations produced by Llama include hallucinations, generating information that does not exist. In contrast, Figure 10 presents the augmentations for the subsequent dialogue turn using the same models. Notably, Claude-Sonnet maintains consistency across both turns, suggesting its stable performance throughout all experiments. While Mistral model tend to be less stable as it included attributes that are not in the dialogue.

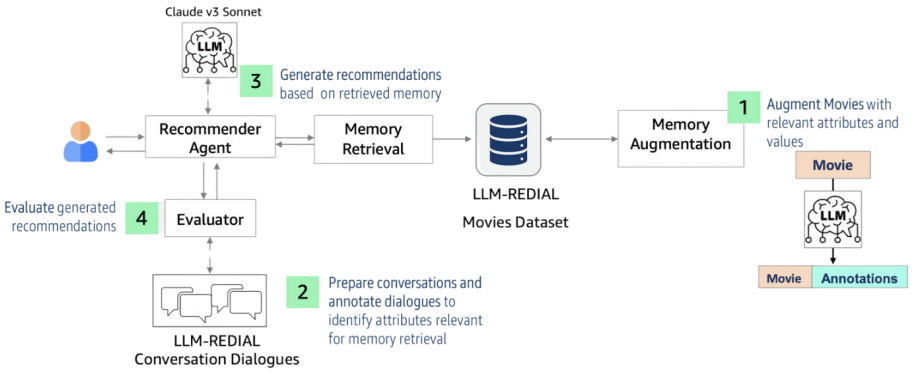

Figure 8: Evaluation Framework for Conversation Recommendation Task.

	Claude-Sonnet	Llama	Mistral
Dialogue Turn	Jon said, "Hey Gina! Good to see you too. Lost my job as a banker yesterday, so I'm gonna take a shot at starting my own business."		
Sample Generated Annotations	**[person]**<Jon> **[job_status]**<lost job> **[former_job]**<banker> **[intent]**<starting own business>	Gina replied, "Sorry to hear that, Jon. What kind of business are you thinking of starting Jon said, "I'm thinking of opening a bakery. I've always been passionate about baking and I think it could be a great opportunity Gina asked, "Do you have any experience running a business Jon replied, "No, but I've got a solid business plan and I'm willing to put in the hard work to make it happen **[Speaker]**<Jon> **[JobLossReason]**<Lost job as a banker> **[NewBusinessIdea]**<Starting a bakery> **[Passion]**<Baking> **[BusinessExperience]**<No> **[BusinessPlan]**<Solid business plan> **[HelpOffered]**<Gina offered to help> **[HelpNeeded]**<Finding a location> **[Location]**<Downtown> **[MeetingTime]**<Tomorrow>	**[person]**<Jon> **[emotion]**<happy> **[occupation]**<banker> **[current_occupation]**<unemployed> **[intention]**<start_business>

Figure 9: Augmentation generated on a Turn-level for a sample dialogue turn from the LoCoMo dataset using Claude-3-Sonnet, Llama v3 and Mistral v1 models.

	Claude-Sonnet	Llama	Mistral
Following Turn	Gina said, "Sorry about your job Jon, but starting your own business sounds awesome! Unfortunately, I also lost my job at Door Dash this month. What business are you thinking of?"		
Sample Generated Annotations	**[person]**<Gina> **[job_status]**<lost_job> **[former_employer]**<Door_Dash> **[person]**<Jon> **[job_status]**<lost_job> **[intent]**<starting_own_business>	Gina said, "Sorry to hear that, Jon. What kind of business are you thinking of starting?"Jon said, "I'm thinking of starting a small consulting firm. I've got a lot of experience in finance and I think I can make a go of it."Gina said, "That sounds great! What kind of consulting firm are you thinking of starting?"Jon said, "I'm thinking of starting a boutique consulting firm that focuses on finance and strategy. **[Speaker]**<Jon> **[JobLossReason]**<Lost job as a banker> **[NewBusinessIdea]**<Starting a bakery> **[Passion]**<Baking> **[BusinessExperience]**<No> **[BusinessPlan]**<Solid business plan> **[HelpOffered]**<Gina offered to help> **[HelpNeeded]**<Finding a location> **[Location]**<Downtown> **[MeetingTime]**<Tomorrow>	**[person.Gina]<name>:** Gina **[person.Jon]<name>:** Jon **[lost_job.Gina]<job_lost>:** Door Dash **[lost_job.Jon]<job_lost>: N/A** **[business_idea]<idea>:** N/A **[emotion.Gina]<emotion>:** sorry **[emotion.Gina]**<emotion>: awesome **[question]<question>**: What business are you thinking of? **[topic]<topic>:** job loss, starting a business.

Figure 10: Augmentations generated for the turn following the turn in Figure 9 using Claude-3-Sonnet, Llama v3 and Mistral v1 models. Hallucinations are presented in red.

Question Augmentation
Given the following question, determine what are the main inquiry attribute to look for and the person the question is for. Respond in the format: Person:[names]Attributes:[].
Basic Augmentation
You are an expert annotator who generates the most relevant attributes in a conversation. Given the conversation below, identify the key attributes and their values on a turn by turn level. Attributes should be specific with most relevant values only. Don't include speaker name. Include value information that you find relevant and their names if mentioned. Each dialogue turn contains a dialogue id between [ ]. Make sure to include the dialogue the attributes and values are extracted form. Important: Respond only in the format [{speaker name:[Dialog id]:[attribute]<value>}]. Dialogue Turn:{}
Priority Augmentation
You are an expert dialogue annotator, given the following dialogue turn generate a list of attributes and values for relevant information in the text. Generate the annotations in the format: [attribute]<value>where attribute is the attribute name and value is its corresponding value from the text. and values for relevant information in this dialogue turn with respect to each person. Be concise and direct. Include person name as an attribute and value pair. Please make sure you read and understand these instructions carefully. 1- Identify the key attributes in the dialogue turn and their corresponding values. 2- Arrange attributes descendingly with respect to relevance from left to right. 3- Generate the sorted annotations list in the format: [attribute]<value>where attribute is the attribute name and value is its corresponding value from the text. 4- Skip all attributes with none vales Important: YOU MUST put attribute name is between [ ] and value between <>. Only return a list of [attribute]<value>nothing else. Dialogue Turn: {}

Table 8: Prompts used in Question Answering for generating augmentations for questions. Also, augmentations for conversations, utilizing both basic and priority augmentations.

Basic Augmentation
For the following movie identify the most important attributes independently. Determine all attributes that describe the movie based on your knowledge of this movie. Choose attribute names that are common characteristics of movies in general. Respond in the following format: [attribute]<value of attribute>. The Movie is: {}
Priority Augmentation
You are a movie annotation expert tasked with analyzing movies and generating key-attribute pairs. For the following movie identify the most important. Determine all attribute that describe the movie based on your knowledge of this movie. Choose attribute names that are common characteristics of movies in general. Respond in the following format: [attribute]<value of attribute>. *Sort attributes from left to right based on their relevance.* The Movie is:{}
Dialogue Augmentation
Identify the key attributes that best describe the movie the user wants for recommendation in the dialogue. These attributes should encompass movie features that are relevant to the user sorted descendingly with respect to user interest. Respond in the format: [attribute]<value>.

Table 9: Prompts used in Conversational Recommendation for recommending Movies utilizing both basic and priority augmentations.

Dialogue Augmentation
Given the following attributes and values that annotate a dialogue for every speaker in the format [attribute]<value>, generate a summary for the event attributes only to describe the main and important events represented in these annotations. Refrain from mentioning any minimal event. Include any event-related details and speaker. Format: a bullet paragraph for major life events for every speaker with no special characters. Don't include anything else in your response or extra text or lines. Don't include bullets. Input annotations: {}

Table 10: Prompt used in Event Summarization to augment dialogues