LLM-Based Behavior Prediction for Social Media Users with Continuous Memory

Kun Li[1,2] **,Chengwei Dai**[1,2]**, Wei Zhou**[1,2*] **Songlin Hu**[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
`{likun2, daichengwei, zhouwei, husonglin}@iie.ac.cn`

Abstract

Large language models (LLMs) have demonstrated strong capabilities in simulating social roles and generating human-like behaviors. However, their effectiveness in predicting real-world user behavior under continuous memory accumulation remains largely unexplored. Most existing studies focus on short-term interactions or static personas, neglecting the dynamic nature of users' historical experiences in social media environments. To address this gap, we introduce FineRob, a novel dataset for fine-grained behavior prediction of social media users, which includes long-term memory traces from 1,866 users across three platforms. Each behavior is decomposed into three elements: object, type, and content, resulting in 78.6k QA records.We identify that as memory accumulates, prediction accuracy drops significantly due to the model's difficulty in accessing detailed historical information. We further propose the OM-CoT fine-tuning framework to enhance the model's ability to process and utilize long-term memory. Experimental results show that our method effectively reduces the performance degradation caused by memory growth, improving fine-grained behavior prediction. [1].

1 Introduction

Large language models (LLMs) have attracted significant attention for their ability to engage in role-playing. These models can be guided by predefined role profiles to generate conversations that align with a character's speaking style (Zhou et al., 2023), knowledge (Lu et al., 2024), and personality traits (Chen et al., 2023). Recently, numerous agent frameworks have been introduced to extend LLMs' capabilities beyond simple dialogue generation(Wang et al., 2023b; DBL, 2024; Wang et al., 2024). However, a critical aspect of human-like intelligence is the ability to maintain consistency and evolve based on continuous experiences, a process heavily reliant on memory. While memory is a core component for social agents, the impact of its continuous growth on behavioral realism remains under-explored.

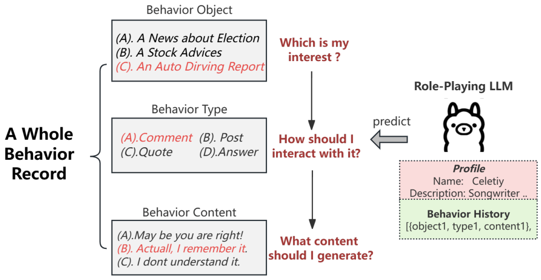

Behavior Object
(A). A News about Election
(B). A Stock Advices
(C). An Auto Dirving Report
Which is my interest ?
Behavior Type
(A).Comment
(B). Post
(C).Quote
(D).Answer
How should I interact with it?
A Whole Behavior Record
Role-Playing LLM
predict
Profile
Name: Celetiy
Description: Songwriter ..
Behavior History
[{object1, type1, content1},
Behavior Content
(A).May be you are right!
(B). Actuall, I remember it.
(C). I dont understand it.
What content should I generate?

Figure 1: An example of FineRob, requires LLM to predict behavior choices that align with a role's profile and historical data. We decompose a complete behavior record into three fine-grained components: selecting the recipient of the action, determining the action type, and specifying the behavior details.

While increasing memory capacity is expected to enhance behavioral coherence and realism, our preliminary experiments reveal a non-linear relationship: as memory accumulates, behavior prediction accuracy first stabilizes and then declines as shown in Figure 2. This suggests that excessive memory may hinder, rather than help, the agent's reasoning process. We hypothesize that this degradation stems from the model's inability to effectively retrieve and utilize detailed historical information when memory becomes too large. Traditional memory compression techniques, which aim to remove redundancy, fail to preserve the fine-grained details necessary for accurate behavior prediction.

To systematically investigate this challenge, we first break down complex user behaviors into three fine-grained elements: the **object** (the target of the action), the **type** (the nature of the action), and the

---

[*] Wei Zhou is the corresponding author.

[1] Code and dataset are available at `https://huggingface.co/datasets/Linkseed/FineRob`

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 459–474
December 20-24, 2025 ©2025 Association for Computational Linguistics

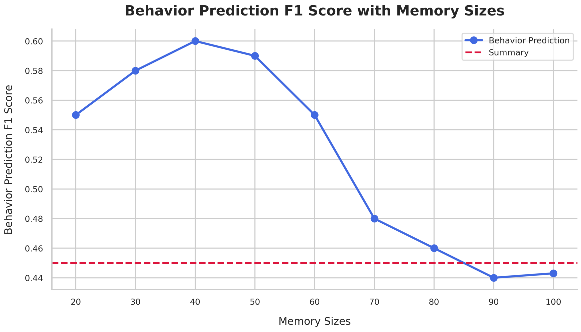


Figure 2: The impact of memory accumulation on behavior prediction accuracy (F1 score) on the Twitter user of Finerob dataset. Performance initially improves with more historical context, peaking at around 40 memory entries. However, further memory accumulation leads to a decline in accuracy, illustrating the challenge of information overload where excessive historical data can impair model performance.

content (the specific substance of the action). This decomposition allows for a precise evaluation of the model's simulation capabilities. For example, when a user comments on a post, they first choose a target (e.g., a post or article from their feed), then decide on the type of action (e.g., comment, like, or share), and finally generate content that aligns with their persona, as illustrated in Figure 1. We then introduce the **FineRob** (**Fine**-Grained **Ro**le **B**ehavior) dataset, a new benchmark designed for this purpose. FineRob contains long-term memory traces from 1,866 real users on Twitter, Reddit, and Zhihu, comprising 78.6k fine-grained behavioral prediction records. Each user profile is supported by a substantial history (averaging 126 records), making it ideal for studying the effects of continuous memory.

Using the FineRob dataset to evaluate nine mainstream LLMs, our analysis of their Chain of Thought (CoT) processes uncovered two dominant reasoning patterns: "**role stereotype-based reasoning**" and "**observation and memory-based reasoning**." The former relies heavily on static profile information, leading to less accurate predictions as memory grows. In contrast, the latter, favored by more advanced models like GPT-4o, achieves higher accuracy by comparing current observations with specific past behaviors. This finding reinforces our hypothesis that the ability to effectively leverage detailed memories is crucial.

Building on this insight, we propose **OM-CoT**, a novel fine-tuning framework designed to mitigate the performance degradation caused by memory growth. The OM-CoT method explicitly integrates **O**bservation Analysis and **M**emory Analysis into the model's reasoning chain. This structured approach enhances the model's ability to connect current environmental cues with relevant, detailed historical memories, thereby improving the accuracy of fine-grained behavior prediction.

The contributions of this study can be summarized as follows:

- **We introduce the FineRob dataset**, a realistic, multilingual benchmark with 78.6k QA records from 1,866 real social media users, specifically designed to evaluate LLMs' ability to simulate role-specific behaviors under conditions of continuous memory growth.
- **We conduct a comprehensive analysis** of how continuous memory growth impacts the behavioral simulation accuracy of nine mainstream LLMs, identifying the failure to access detailed memory as a key cause of performance degradation.
- **We propose OM-CoT**, a novel fine-tuning method that explicitly integrates observation and memory analysis into the reasoning process. Our experiments demonstrate that this method significantly mitigates performance decline and improves prediction accuracy across all three fine-grained behavioral elements.

2 Related Work

Recently, numerous LLMs have been designed for conversational applications and have already found commercial uses like Character.ai[2].Researchers collect a wide range of dialogue datasets to support the study, including the data from real-life individuals (Gao et al., 2023; Dinan et al., 2019) or fictional characters from novels (Chen et al., 2023; Ahn et al., 2024).Researchers have recognized that LLMs are not only adept at mimicking conversation but also capable of simulating complex behaviors (Park et al., 2023; Wang et al., 2023b; Chen et al., 2024; Wang et al., 2024). For instance, (Wang et al., 2023a) demonstrated that LLMs can mimic real users' preferences to movie recommendations. (Ebrat and Rueda, 2024) explored the use of reinforcement learning algorithms to optimize recommendation systems based on user feedback simu-

---

[2] `https://character.ai/`

Dataset	Source	Size	Usage	Language	Type	isReal?	Context
PersonalDialog(Zheng et al., 2019)	Weibo	20.8M	Train&Test	ZH	Dialogue	✓	3 turns conversations
Ditto(Lu et al., 2024)	LLM Synthetic	4k	Train&Test	EN,ZH	Dialogue	×	30 turns conversations
LaMP-7(Salemi et al., 2024)	Twitter	12k	Train&Test	EN	Dialogue	✓	1 turn conversation
PIPPA(Gosling et al., 2023)	Character.AI	26k	Train	EN	Dialogue	×	10 turns conversations
ROCStories(Mostafazadeh et al., 2016)	Little Stories	98K	Train&Test	EN	Behavior	×	short paragraph
choices13K(Bourgin et al., 2019)	gamble	13K	Test	EN	Behavior	✓	short paragraph
Life-Choice(Xu et al., 2024)	Novels	1.4K	Test	EN	Behavior	×	long fiction
FineRob	SocialMedias	78.6K	Train&Test	EN,ZH	Behavior	✓	74 turns behaviors

Table 1: A comparison of FineRob with previous datasets based on data source, type, and context length. The table indicates that existing datasets typically provide context in the form of multi-turn conversations or single behavioral scenarios. FineRob is characterized by its long-term, continuous context, containing an average of 74 sequential behavior records per real user.

lated by LLMs. The work by (Xu et al., 2024) is particularly relevant, which introduced the "LIFE-CHOICE" dataset to assess LLMs' ability to make broad, macro-level decisions in fictional contexts. In contrast, our research focus on the LLMs' capacity to simulate fine-grained, micro-level behaviors in real world scenario, offering a more detailed analysis.

3 FineRob Dataset

3.1 Data Collecting

Our primary goal is to investigate how LLMs simulate the behaviors of real internet users by leveraging their past experiences. A key departure from existing role-playing datasets, which often rely on static character profiles, is our focus on collecting long-term, continuous memory traces. We posit that this historical data is crucial for simulating nuanced, evolving behaviors. To this end, we curated a dataset from three major social media platforms: Twitter (now X), Reddit, and the Chinese QA site Zhihu. The dataset is designed to capture extensive behavioral histories, providing a rich foundation for memory-based analysis, as illustrated in Figure 3. A detailed comparison with other datasets is provided in Table 1.

To ensure the relevance and activity of our subjects, we selected users from trending topics and popular communities on each platform.

- On Reddit, we identified active participants in posts from the top 20 communities and used the PRAW API[3] to automatically collect their complete public timelines.
- On Twitter, we utilized the official API[4] to access and gather data from users engaged in trending discussions.
- On Zhihu, we targeted users who frequently post questions or answers.

We applied minimal selection criteria: users needed at least 60 accessible historical behaviors to provide a sufficient memory base, while those with excessive activity were excluded to avoid accounts potentially managed by teams. Crucially, to preserve the authenticity of real-world interactions, the collected behavioral data is entirely organic. We performed no manual filtering or content moderation on the user timelines, ensuring that the dataset reflects the genuine complexity and unpredictability of human behavior online.

3.2 Fine-Grained Behavior and Choice Construction

We converted the raw user timelines into a multiple-choice QA format for fine-grained behavior prediction. Each behavior record was deconstructed into three key elements: the **object** (the recipient of the behavior), the **type** (the nature of the action), and the **content** (the specific details). This process is detailed in the middle section of Figure 3.

A significant challenge was constructing realistic and challenging multiple-choice options. The behavior type options were straightforward, derived from predefined platform actions like "Post," "Comment," or "Like" (see Table 6 for a full list).For the more complex object and content elements, we developed a sophisticated method to create plausible distractors. This process also included a "None of the above" option to test the model's ability to reject unsuitable choices, not just identify the best match.

- **Constructing Plausible Distractors:** For a standard question where a correct answer ex-

---

[3] `https://praw.readthedocs.io/en/stable/`

[4] `https://developer.x.com/en/docs/x-api`

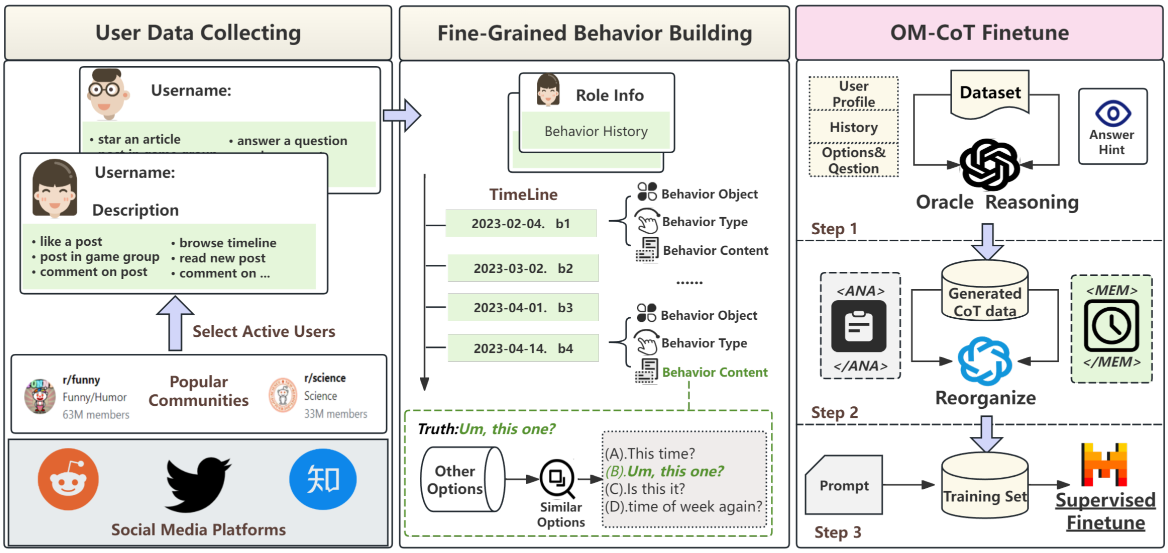


Figure 3: Overview of our work, The left and middle sections of the figure illustrate the process of constructing the FineRob dataset. The right section shows how OM-COT-FineTune training details, including data augmentation, reorganize with special tokens and SFT training.

ists among the options, we first created a candidate pool of objects or content from the user's active communities and timeframes. We then used sentence embeddings[5] to measure the semantic similarity between these candidates and the ground-truth answer. To maximize the difficulty, we selected three candidates with high similarity scores to serve as confusing distractors.

- **Constructing "None of the Above" Scenarios:** To further test the model's reasoning, we designed scenarios where the correct answer is "None of the above." For these instances, the provided options are deliberately sourced from contexts irrelevant to the specific moment of action. Specifically, we analyzed a user's typical active hours and then selected plausible-looking objects or content from their **inactive periods** from other users' feeds. While these options might be thematically related to the user's general interests, they are incorrect in the specific context. This forces the model understand deeply of user rather than simply choosing the most semantically similar option.

Finally, we compiled a total of 78.6k behavior element prediction records. The dataset was split into a training set of approximately 61k records and a test set of 17.6k records. To rigorously evaluate the generalization capabilities of LLMs, we ensured that no users overlap between the training and test sets. A detailed example of a data record can be found in Table 7 of the appendix.

4 Methodology

4.1 Preliminary Analysis

Our preliminary experiments reveal a counter-intuitive phenomenon: while an agent's behavioral realism gradually improves and stabilizes as it references more memory within a certain range, its accuracy begins to decline once the memory volume exceeds a threshold, as shown in Figure 2. This indicates that memory accumulation is not monotonically beneficial; excessive storage can interfere with the model's reasoning capabilities. To investigate potential causes, we attempted to use existing memory summarization and compression methods to filter redundant memories, but these failed to mitigate the performance degradation.

We therefore hypothesize that the core problem lies not in the volume of memory, but in the model's difficulty in extracting **critical, fine-grained details** from a vast repository of historical data. This leads to a key question: which memories are most conducive to realistic behavioral decisions? To understand this process, we analyzed the agent's reasoning process, which can be divided into two key parts: (1) analysis of the current environment (ob-

---

[5] https://huggingface.co/moka-ai/m3e-base

	Reddit			Twitter			Zhihu		
	BO	BC	BT	BO	BC	BT	BO	BC	BT
Random	10.32	10.46	6.31	10.45	8.65	11.85	10.14	10.22	3.12
Commercial-LLM									
chatgpt-3.5-turbo-0125	19.99±0.2	19.72±0.1	51.22±0.0	54.15±0.2	37.62±0.0	62.33±0.4	25.09±0.0	33.17±0.0	19.24±0.0
GPT-4-mini	26.49±0.0	23.91±0.0	48.14±0.0	**63.33**±0.0	41.26±0.0	85.99±0.0	31.65±0.0	31.94±0.0	24.04±0.0
GPT-4o	**28.34±1.2**	**24.13±0.0**	**58.14±0.0**	53.70±0.0	**52.90±0.0**	**86.97±0.0**	**36.34±0.1**	**41.49±0.0**	**26.35±0.0**
Open-LLM									
Mistral-7b-v2.0	25.95±2.0	19.62±1.4	**22.92±0.7**	19.99±2.4	27.04±6.5	62.92±0.9	21.56±1.2	19.53±1.6	10.38±0.2
LLama3-8b	10.80±0.5	11.8±0.5	13.6±0.6	22.72±0.9	16.23±3.5	52.90±0.7	21.70±0.9	12.66±1.2	10.75±0.2
Solar-10.7b	**27.31±1.6**	**25.31±2.3**	20.52±0.8	**53.40±1.6**	27.72±7.6	71.12±1.4	19.48±1.6	15.00±2.1	10.66±0.5
Baichuan-13b	27.13±1.4	21.98±1.1	21.16±0.8	16.22±0.6	12.64±0.3	37.51±1.1	**31.27±0.8**	24.62±1.5	15.97±0.5
yi-34b	25.94±1.9	24.94±1.8	18.26±1.6	51.14±0.7	32.74±2.9	**73.98±0.8**	33.93±1.1	**26.84±1.2**	16.15±0.4
qwen-2-72b	12.37±0.7	11.19±0.7	15.68±2.5	52.44±6.7	**45.39±5.5**	**77.65±6.1**	30.92±0.6	22.35±1.4	**16.80±0.3**

Table 2: F1-scores of nine LLMs under a zero-shot setting, where BO, BC, and BT represent Behavior Object, Behavior Content, and Behavior Type, respectively. The average and standard deviation were recorded over multiple runs. The best and second-best results from Commercial-LLM and Open-LLM are highlighted using **bold** and underline formatting.

servation) and memory retrieval, and (2) behavior decision-making based on the retrieved memory (as illustrated in Figure 5). Our results show a strong positive correlation between the agent's accuracy and the degree of granularity matching between the retrieved memory and the current observation (Figure 6). When the memory details align well with the fine-grained information in the current context, the agent's decisions are more human-like. Conversely, when memories are too abstract or lack key details, decisions become less realistic. This finding confirms our hypothesis that a **mismatch in memory granularity** is a key factor behind the decline in behavioral realism. The detail analysis and figures can be found in Appendix A.1.

4.2 OM-CoT

Based on the foregoing analysis, we propose a simple yet effective method named OM-CoT (Observation and Memory-based Chain of Thought) Finetune. Traditional memory management methods often focus on removing redundancy, but risk losing the very details that are crucial for reasoning. Our approach, in contrast, is designed to enhance the model's ability to synthesize information from current observations and historical memory through a structured reasoning process. The core idea of OM-CoT is to explicitly integrate Observation Analysis and Memory Analysis into the Chain of Thought (CoT) process. This encourages a reasoning pattern based on linking observations to specific, detailed memories, thereby ensuring logical consistency and behavioral realism. The overall method, illustrated in Figure 3, consists of the following three steps.

Step 1: Oracle CoT Generation. To construct a high-quality dataset for our structured reasoning format, we first use a powerful pretrained language model (M_{oracle}) to generate the CoT logic. To prevent the propagation of errors from flawed free-form generation, we employ an *oracle* setting. In this setup, the correct answer (y^*) is provided in the input prompt alongside the sample data (x), ensuring the model references the ground-truth behavior during its reasoning process. The Oracle CoT generation is defined as:

$$C^* = M_{\text{oracle}}(x, y^*) \tag{1}$$

Here, C^* represents the high-quality CoT generated by the model. We carefully design the prompt to ensure that the CoT explains the reasoning path without directly revealing the final answer, thus guaranteeing the quality of the data for subsequent fine-tuning.

Step 2: Reorganizing CoT with Special Tokens. Next, we introduce a special token mechanism to structurally reorganize the generated CoT. This allows the model to explicitly distinguish between reasoning based on the current environment and details recalled from memory. We define two pairs of special tokens:

- `<ANA>` and `</ANA>` wrap the analysis based on the agent's current observation.
- `<MEM>` and `</MEM>` wrap the reproduction of details from the agent's historical memory.

We use a smaller large language model (M_{reorg}) to perform this reorganization on the generated CoT

(C^*):

$$C^{\text{reorg}} = M_{\text{reorg}}(C^*) \tag{2}$$

The resulting reorganized CoT, C^{reorg}, follows a structured format. This token-based structure ensures that the model learns to correctly identify and utilize information from different sources, enhancing the transparency and controllability of the reasoning process. A complete example is provided in Tabel 8.

Step 3: Supervised Fine-Tuning. In the final stage, we use the reorganized CoT dataset to perform Supervised Fine-Tuning on our target LLM. This teaches the model to effectively use the special tokens to structure its reasoning process. Given a training sample $(x, H, C^{\text{reorg}}, y)$, where H is the user history, we minimize the standard language model loss:

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|C^{\text{reorg}}|} \log P\left(c_t^{\text{reorg}} \mid x, H, c_{<t}^{\text{reorg}}\right) \tag{3}$$

where c_t^{reorg} is the t-th token of the reorganized CoT sequence. Through this process, the model not only optimizes its language modeling task but also learns to leverage the special tokens and system prompts to control its CoT reasoning flow. We constructed a dataset of 60,000 instruction-tuning examples for this purpose, which is publicly available in our codebase.

In summary, compared to traditional memory management, OM-CoT focuses on maintaining fine-grained memory matching during reasoning.

5 Experiment

5.1 Settings

Our main experiments are divided into two parts. The first part establishes a baseline for behavior prediction with continuous memory by evaluating a wide range of current LLMs on the FineRob dataset. The models tested include mainstream commercial models, general-purpose open-source models, and open-source models specifically fine-tuned for role-playing. The results is shown in Table 2. The second part of our experiments analyzes the effectiveness of various memory utilization strategies. We compare several methods for providing historical context, including using **all** memories, using only **recent** memory (the last 30 records), and retrieval-based approaches using **BM25** and the **BGE** embedding model to find similar memories. We then compare a standard **CoT)** method against our proposed **OM-CoT** method. The results is shown in Table 3.To mitigate the effects of randomness inherent in large models, each experiment was run three times, and we report the mean and variance of the results. For a detailed description of all configurations, please refer to the settings section in Appendix A.2. We adopt the F1-score of behavior selection as the evaluation metric.

5.2 Main Result

Bigger Models Are Not Always Better We observe that larger open-source models do not consistently outperform smaller alternatives, which is unexpected. For example, Qwen-2-72B, one of the top open-source models, shows strong performance on multilingual tasks, especially with the Zhihu and Twitter datasets. However, its results on the Reddit dataset fell significantly below expectations. In contrast, smaller models like Solar-10.7B and Mistral-7B, despite having fewer parameters, deliver more balanced and competitive outcomes across a broader range of tasks.

OM-CoT Fine-Tuning Enhances Behavior Simulation Performance The experimental results, presented in Table 3, first highlight the challenges of utilizing continuous memory in behavior prediction. A key observation from our baseline analysis is that directly providing the model with the entire memory history often results in poor performance. This supports our initial hypothesis that an overabundance of information can impair the model's ability to focus on relevant details. Furthermore, the analysis reveals that conventional retrieval methods, such as BM25 and BGE-based retrieval, do not offer a significant improvement over simply using the most recent memory records. This suggests that merely finding semantically similar past behaviors is insufficient; the model still struggles to effectively integrate these retrieved memories into its reasoning process. In contrast to the limited effectiveness of these baseline strategies, our OM-CoT fine-tuning method yields substantial and consistent improvements. As shown in Table 3, applying OM-CoT to the Mistral-7B model enhances performance across all nine sub-tasks, achieving an average F1 score increase of approximately 4.5%. Similarly, the Solar-10.7B model demonstrates gains in seven out of nine tasks, including a notable 9.8% improvement on the Reddit Behavior Object prediction task. A qualitative ex-

	Reddit			Twitter			Zhihu		
	BO	BC	BT	BO	BC	BT	BO	BC	BT
Mistral-7b-Instruct									
ALL	19.99±0.0	14.18±0.0	24.01±0.0	38.80±0.0	22.47±0.0	56.20±0.0	20.32±0.4	23.51±1.2	14.51±0.2
Recent	25.95±2.0	19.62±1.4	22.92±0.7	19.99±2.4	27.04±6.5	62.92±0.9	21.56±1.2	19.53±1.6	10.38±0.2
RAG-BM25	25.33±0.1	19.28±0.0	23.02±0.0	18.86±0.0	28.11±0.0	64.52±0.0	22.73±0.1	16.66±0.2	13.44±0.0
RAG-BGE	24.23±0.0	20.47±0.1	21.53±0.0	17.32±0.0	27.56±0.0	65.35±0.0	23.82±0.0	17.89±0.1	15.21±0.0
CoT-FT	31.56±0.0	19.46±0.1	31.40±0.0	55.76±0.0	55.56±0.0	84.14±0.0	29.30±0.4	16.61±0.8	17.77±0.0
OM-CoT-FT(ours)	**34.58±0.2**	**21.27±0.2**	**45.12±0.7**	**64.38±0.0**	**56.25±0.0**	**88.12±0.0**	**33.19±0.7**	**28.66±0.5**	**21.54±0.1**
Solar-10.7b-Instruct									
ALL	18.31±0.0	16.44±0.0	21.30±0.0	38.84±0.0	22.74±0.0	56.20±0.0	19.94±0.1	11.26±0.0	12.46±0.2
Recent	27.31±1.6	**25.31±2.3**	20.52±0.8	53.40±1.6	27.72±7.6	71.12±1.4	19.48±1.6	15.00±2.1	10.66±0.5
RAG-BM25	25.45±0.2	22.91±0.0	18.17±0.0	48.93±0.0	25.41±0.0	68.32±0.0	18.66±0.1	14.16±0.1	10.31±0.1
RAG-BGE	26.23±0.0	22.15±0.1	20.32±0.1	50.41±0.0	26.41±0.0	70.83±0.0	20.12±0.1	19.31±0.2	10.67±0.0
CoT-FT	28.30±0.0	21.20±0.1	45.33±0.0	62.27±0.0	**58.86±0.0**	74.04±0.0	24.63±0.3	34.21±0.8	16.18±0.4
OM-CoT-FT(ours)	**38.16±0.4**	23.23±0.0	**47.44±0.1**	**73.53±0.0**	49.04±0.0	**87.29±0.0**	**28.45±0.1**	**36.65±0.4**	**19.53±0.0**

Table 3: F1-scores of OM-COT-FT and other baselines, where BO, BC, and BT represent Behavior Object, Behavior Content, and Behavior Type, respectively.

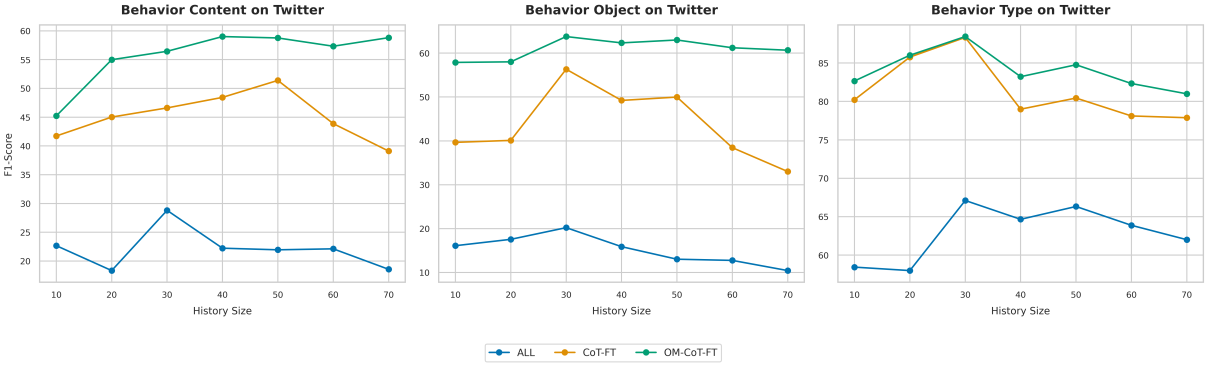


Figure 4: The relationship between input historical behavior size and the accuracy of simulating fine-grained behavior elements. The figure presents the results of three methods on the Twitter dataset.

ample of the improved reasoning process can be found in Table 9. Crucially, both standard CoT fine-tuning (Std-CoT-FT) and our OM-CoT-FT consistently outperform the non-fine-tuned models. This result is particularly significant because the training and test sets contain entirely different users. It indicates that the models are not merely memorizing user-specific patterns but are learning a generalizable reasoning framework that can be effectively applied to new, unseen user roles.

LLMs Struggle with Short-Behavior Tasks A notable result emerge from the Reddit dataset, particularly in the behavior content prediction task, where fine-tuning methods failed to yield performance improvements. Upon further investigation, we find that Reddit content is often brief and lacks clear indicators of user characteristics. This suggests that current language models still struggle to differentiate subtle variations in tone and punctuation within behaviors (e.g., "Good work" vs. "Pretty Nice!!").

5.3 Discussion

In this subsection, we will conduct ablation studies to further analyze the key factors that influence behavior simulation performance. Specifically, we seek to address the following three research questions.

		BO	**BC**	**BT**
Mistral-7b (Zero-Shot)	ALL	19.99	27.04	62.92
	w/o userinfo	17.59$_{-2.4}$	17.63$_{-9.4}$	53.81$_{-9.1}$
	w/o interest	19.75$_{-0.2}$	24.04$_{-3.0}$	54.60$_{-8.3}$
	w/o history	11.99$_{-6.0}$	21.14 $_{-5.9}$	26.21$_{-36.7}$
Mistrail-7b (OM-Cot-FT)	ALL	64.38	56.25	88.12
	w/o userinfo	64.61$_{+0.3}$	46.19$_{-10.1}$	86.60$_{+0.5}$
	w/o interest	65.49$_{+0.1}$	50.48$_{-5.8}$	86.06$_{-2.1}$
	w/o history	39.60$_{-24.7}$	43.12$_{-13.17}$	43.42$_{-44.8}$

Table 4: Ablation study on different prompt components to explore how various aspects influence the simulation of fine-grained behavior elements.

RQ1: Which part of the prompt is the most important? Behavior simulation prompt includes role's basic information, interests, and past behaviors. To assess the importance of each component, we conduct ablation experiments by removing individual parts from the input prompts.Table 4 shows the results on the Twitter dataset, demonstrating how these components affect model performance. As highlighted in our preliminary experiments, role history is the most influential, especially for OM-CoT fine-tuned models that are trained to analyze historical behaviors. Removing role history leads to a notable performance drop. On the other hand, the effect of basic information and interests varies across different behavior elements. For example, excluding basic info and interests has minimal impact on predicting behavior object and type, but they are useful for accurately predicting behavior content.Notably, while OM-CoT emphasizes observation and memory-based reasoning, it still integrates character profile analysis within the CoT process, which relies on role information in the input prompt.

RQ2: How does the continuous accumulation of memory impact behavior prediction accuracy? A central question for our research is how prediction accuracy is affected as an agent's memory continuously accumulates over time. To simulate this, we evaluated model performance by progressively increasing the number of historical records provided as input, from the 10 most recent entries up to the user's entire available history (average 74 records).The results, illustrated in Figure 4, reveal a critical challenge for standard models. Contrary to the simple assumption that more data is always better, the performance of baseline methods does not improve monotonically. Instead, their F1 scores typically peak with a moderate amount of history (around 30 50 records) and then exhibit significant fluctuation and decline as more memories are added. This trend strongly supports our hypothesis that without a proper mechanism to handle information overload, increasing memory introduces noise that impairs the model's ability to pinpoint relevant details, thus degrading prediction accuracy.In contrast, models fine-tuned with our **OM-CoT** method demonstrate remarkable robustness. As the figure shows, the performance of the OM-CoT models remains high and stable, even as the volume of historical memory grows.

RQ3:Do both the <ANA> and <MEM> special token work effectively? To investigate this, we conducted ablation experiments by selectively removing content enclosed by the special tokens `<ANA>` (analysis) and `<MEM>` (memory) during the reasoning process. This was done by either adjusting the system prompt or excluding these tokens during the decoding process. The results are shown in Table 5.The experimental findings reveal that removing either `<ANA>` or `<MEM>` leads to a decrease in behavior simulation performance, highlighting the importance of both tokens in the CoT reasoning process. However, the influence of these tokens varies across different sub-tasks. For behavior type prediction, the model relies more on analyzing available candidate options (`<ANA>`), whereas behavior content prediction depends more on recalling and reproducing historical behaviors (`<MEM>`). In the case of behavior object prediction, both observation and memory are equally important. This suggests that each sub-task requires a different balance between reasoning based on current observations and past behavior records to achieve optimal performance.

	BO	**BT**	**BC**
OM-CoT-FT	64.38	56.25	88.12
only `<ANA>`	61.36$_{-3.0}$	46.46$_{-9.8}$	84.37$_{-3.8}$
only `<MEM>`	58.88$_{-5.5}$	55.58$_{-0.7}$	72.25$_{-15.9}$

Table 5: The ablation experiment of two special tokens used in our OM-COT-FT method.The table presents the results using the Mistral-7b-instruct model on the Twitter dataset.

6 Conclusion

In this work, we address the critical challenge of performance degradation in LLM agents as their memory grows. Using our new benchmark, **FineRob**, we identify a "memory granularity mismatch" as the primary cause: models fail to extract key details from vast historical data. To resolve this, we propose **OM-CoT**, a fine-tuning framework that explicitly structures reasoning into "Observation" and "Memory" analysis steps. Comprehensive experiments demonstrate that OM-CoT effectively mitigates the performance decline seen in baseline models, maintaining high and stable prediction accuracy in memory-rich scenarios. This work provides a path toward building more robust LLM agents that can learn from continuous experience without being hindered by it.

Limitations

A limitation lies in the handling of building behavior object task's options. While our method for constructing "None of the above" scenarios is designed to rigorously test contextual reasoning, it relies on a heuristic that introduces a potential limitation. We source these distractor options from a user's typical "inactive periods," assuming any action during these times is contextually invalid. However, human behavior is not always predictable, and a user might occasionally be active outside their established patterns. Consequently, it is possible that a small fraction of these "None of the above" instances are mislabeled, as a real user could have plausibly chosen one of the provided options. Despite this, we believe this heuristic remains a pragmatic and effective approach for evaluating a model's ability to recognize contextual mismatches, which is a key goal of our study.

Ethics Statement

Current LLMs have achieved highly realistic role-playing abilities, sometimes indistinguishable from real human. This development unlocks potential applications in areas like companionship, entertainment, and education, where human-like agents could lead to significant progress. However, these capabilities also pose risks in social networks, such as the spread of propaganda, misinformation, and the proliferation of malicious bots.This study aims to enhance LLM' ability to accurately simulate human roles while also providing insights into identifying behavioral patterns and thinking styles of such LLM-driven social bots.

References

2024. Enhancing role-playing systems through aggressive queries: Evaluation and improvement. *CoRR*, abs/2402.10618.

Jaewoo Ahn, Taehyun Lee, Junyoung Lim, Jin-Hwa Kim, Sangdoo Yun, Hwaran Lee, and Gunhee Kim. 2024. Timechara: Evaluating point-in-time character hallucination of role-playing large language models. In *ACL (Findings)*, pages 3291–3325. Association for Computational Linguistics.

David D. Bourgin, Joshua C. Peterson, Daniel Reichman, Stuart J. Russell, and Thomas L. Griffiths. 2019. Cognitive model priors for predicting human decisions. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5133–5141. PMLR.

Jing Chen, Xinyu Zhu, Cheng Yang, Chufan Shi, Yadong Xi, Yuxiang Zhang, Junjie Wang, Jiashu Pu, Rongsheng Zhang, Yujiu Yang, and Tian Feng. 2024. Hollmwood: Unleashing the creativity of large language models in screenwriting via role playing. *CoRR*, abs/2406.11683.

Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. 2023. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In *EMNLP (Findings)*, pages 8506–8520. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR (Poster)*. OpenReview.net.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Danial Ebrat and Luis Rueda. 2024. Lusifer: Llm-based user simulated feedback environment for online recommender systems. *CoRR*, abs/2405.13362.

Jingsheng Gao, Yixin Lian, Ziyi Zhou, Yuzhuo Fu, and Baoyuan Wang. 2023. Livechat: A large-scale personalized dialogue dataset automatically constructed from live streaming. In *ACL (1)*, pages 15387–15405. Association for Computational Linguistics.

Tear Gosling, Alpin Dale, and Yinhe Zheng. 2023. PIPPA: A partially synthetic conversational dataset. *CoRR*, abs/2308.05884.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. SOLAR 10.7b: Scaling large language models with simple yet effective depth upscaling. *CoRR*, abs/2312.15166.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In *ACL (1)*, pages 7828–7840. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *HLT-NAACL*, pages 839–849. The Association for Computational Linguistics.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST*, pages 2:1–2:22. ACM.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *ACL (1)*, pages 7370–7392. Association for Computational Linguistics.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, et al. 2023a. User behavior simulation with large language model based agents. *arXiv preprint arXiv:2306.02552*.

Xi Wang, Hongliang Dai, Shen Gao, and Piji Li. 2024. Characteristic AI agents via large language models. In *LREC/COLING*, pages 3016–3027. ELRA and ICCL.

Zhilin Wang, Yu-Ying Chiu, and Yu Cheung Chiu. 2023b. Humanoid agents: Platform for simulating human-like generative agents. In *EMNLP (Demos)*, pages 167–176. Association for Computational Linguistics.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqing Dong, and Yanghua Xiao. 2024. Character is destiny: Can large language models simulate persona-driven decisions in role-playing? *CoRR*, abs/2404.12138.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm:

Customizing chinese conversational AI characters with large language models. *CoRR*, abs/2311.16832.

A Appendix

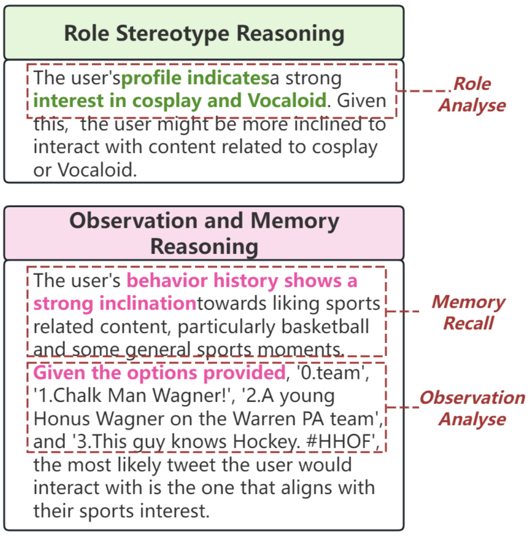
A.1 Behavior Reasoning Analysis

To understand why LLMs struggle with continuous memory, we conducted a detailed analysis of the Chain-of-Thought (CoT) reasoning processes they employ during behavior simulation. Our analysis identified two distinct and competing reasoning patterns, which we term "Role Stereotype-based Reasoning" and "Observation and Memory-based Reasoning."

Qualitative Analysis of Reasoning Patterns. As illustrated in Figure 5, these two patterns represent fundamentally different approaches to the prediction task.

- **Role Stereotype-based Reasoning:** This pattern relies heavily on analyzing the static user profile provided in the prompt. The model deduces behavior based on general personality traits, interests, or demographic information. For instance, it might reason, "This user is described as a tech enthusiast, so they are likely to comment on a post about new gadgets." While seemingly logical, this approach often fails because it ignores the specific context of the current situation and the nuances of the user's actual past behaviors. It leads to generic and often inaccurate predictions, especially as the memory of specific past actions grows.

- **Observation and Memory-based Reasoning:** This superior pattern involves a more dynamic, two-step process. First, the model analyzes the fine-grained details of the currently observed options. Second, it actively searches through the provided historical memory to find analogous situations or relevant past behaviors. For example, it might reason, "The user is presented with an option to comment on a political article. I see in their memory that they have consistently avoided political discussions in the past, despite commenting on other news. Therefore, they are unlikely to choose this option." This method, which connects current observations to specific historical evidence, results in more accurate and human-like simulations.



The user'sprofile indicatesa strong interest in cosplay and Vocaloid. Given this, the user might be more inclined to interact with content related to cosplay or Vocaloid.

The user's behavior history shows a strong inclinationtowards liking sports related content, particularly basketball and some general sports moments. Given the options provided, '0.team', '1.Chalk Man Wagner!', '2.A young Honus Wagner on the Warren PA team', and '3.This guy knows Hockey. #HHOF', the most likely tweet the user would interact with is the one that aligns with their sports interest.

Figure 5: Two typical patterns of COT reasoning for behavior simulation. The "Role Stereotype" pattern focus on role analysis. The "Observation and Memory" pattern predicts future behavior by considering the relationship between the character's history and observed options.

Quantitative Evidence for the Memory Granularity Mismatch. To quantitatively validate these qualitative observations, we analyzed the correlation between the reasoning text and different parts of the input prompt. Specifically, we measured the semantic similarity between the generated CoT and three components: the user's role profile, the observed options, and the historical memory.

The results, shown in Figure 6, provide strong evidence for our "memory granularity mismatch" hypothesis. We found that a higher similarity between the CoT and the **observed options** or **historical memory** is positively correlated with higher prediction accuracy (F1-score). This indicates that when the model successfully focuses on and reasons about the specific details of the current context and relevant past events, its performance improves.

Conversely, a high similarity to the **role profile** does not guarantee better performance. This confirms that simply reiterating general stereotypes from the user description is an ineffective strategy. The core challenge, therefore, is the model's inability to consistently access and leverage the fine-grained details within a large and growing memory store. This difficulty is precisely what leads to the performance degradation observed in our experiments and what the OM-CoT framework is designed to overcome.

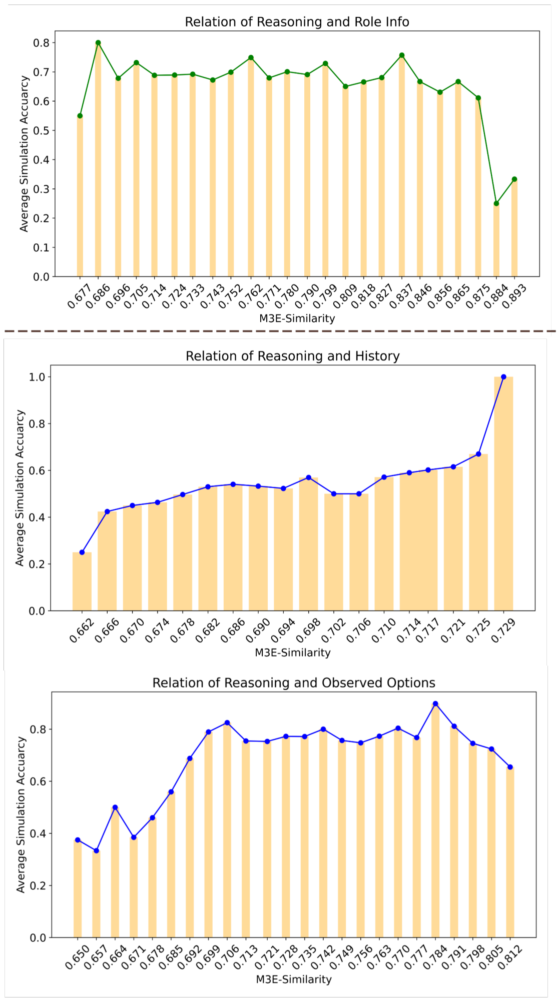


Figure 6: Analysis of simulation accuracy changes across different similarity levels between reasoning and various parts of the prompt. The results are generated using ChatGPT-3.5-turbo-0125 on the Twitter test set, with the average F1-score calculated across three behavior element tasks.

A.2 Experiment Settings

Models We evaluate a total of nine large language models on FineRob. This includes three commercial LLMs, ChatGPT-3.5-turbo-0125/GPT-4o-mini/GPT-4o[6],as well as six open-source LLMs:: Mistral-7b-Instruct (Jiang et al., 2023),Llama3-8b-Instruct (Dubey et al., 2024), Solar-10.7b-Instruct (Kim et al., 2023), Yi-1.5-34B-Chat (Young et al., 2024), Baichuan2-13B-Chat (Yang et al., 2023),and Qwen2-72B-instruct (Yang et al., 2024). The last two models is specialize for Chinese context.

Baselines We conduct extended experiments on the Mistral-7b-Instruct and Solar-10.7b-Instruct models using four baseline methods: including using **all** memories, using only **recent** memory (the last 30 records), and retrieval-based approaches using **BM25** and the **BGE** embedding model to find similar memories. The standard **CoT fine-tuning** method uses un-reorganized CoT data without special tokens. By comparing these approaches, we aim to evaluate how different training and prompting methods affect LLM behavior simulation performance.

Prompts The prompts were similarly structured across all baseline methods and consisted of four main parts: (1) a task description instructing the model to predict three behavior elements while role-playing a specific character, (2) the role's profile, which included username, self-description, and areas of interest, (3) behavior history, detailing the target, type, content, and timing of past behaviors, and (4) method-specific instructions and output format requirements. For example, in OM-CoT, the model was instructed to use the `<ANA>` and `<MEM>` tokens for analysis based on observation and memory.

Implementation Details We utilize LoRA (Hu et al., 2022) for efficient parameter fine-tuning, setting α to 1.0 and β to 0.025. All training are conducted with fp16 mixed-precision on 4 × A100 GPUs over 10 epochs with LLama-factory[7](Zheng et al., 2024). For inference, we use vLLM[8] to accelerate the process, employing sampling decoding with a temperature of 0.1.The F1 score serves as the evaluation metric across all experiments. To mitigate the impact of randomness inherent in LLMs, we run three trials and compute the mean and standard deviation, ensuring more reliable results.

A.3 Examples

This section provides concrete examples of our dataset, training data, and model outputs to complement the descriptions in the main paper. These cases illustrate the structure of our data and the tangible output of the OM-CoT method.Table 6 lists the primary behavior types collected from each social media platform. These actions form the ground truth for the "behavior type" prediction sub-task. The set of possible actions is finite and platform-dependent, providing a clear classification challenge for the models.

---

[6] https://platform.openai.com/docs/models

[7] https://github.com/hiyouga/LLaMA-Factory

[8] https://github.com/vllm-project/vllm

Name	Description	Need Target?	Need Content?
	Reddit		
comment	comment to post or other comment on reddit	✓	✓
post	create a new post on subreddit	×	✓
	Twitter		
replied to	replied to other tweets or comments	✓	✓
post	create a new tweet	×	✓
like	approval or support for a tweet without sharing.	✓	×
quoted	Adds your comment to someone else's shared tweet.	✓	✓
retweet	shares someone else's tweet with your followers unchanged.	✓	×
	Zhihu		
new question	Ask a question to seek answers	✓	×
answer	Answer a question to share knowledge	✓	✓
opinion	Post a thought to share opinions	×	✓
post article	Post a new article	×	✓
update question	Update a question	✓	✓
agree answer	Agreed with an answer by upvoting to show support or approval.	✓	×
follow question	Follow a question to receive updates or answers about it.	✓	×
agree article	Agreed with an article by upvoting to show support or approval.	✓	×
bookmark article	Saved an article to bookmark it for later reference or reading	✓	×
bookmark answer	Saved an answer to bookmark it for later reference or reading	✓	×
approve answer	Approve an answer by endorsing its accuracy or helpfulness	✓	×

Table 6: Valid behavior types on three platforms

The foundation of our work is the collection of authentic, long-term user behavior histories. Table 7 presents a snippet of a real user's timeline from Twitter. This raw data is chronological and contains a mix of actions (e.g., *Post*, *Like*, *Retweet*) and content (text, hashtags, links). It is this rich, and often noisy, historical context that the model must learn to navigate. The challenge lies in identifying the meaningful patterns within this stream of interactions.

The final result of our method is a model that can autonomously apply this structured reasoning to new, unseen scenarios. Table 9 displays the output of our fine-tuned Mistral-7B model on a case from the Reddit test set. The model, without any oracle guidance, first analyzes the current options regarding financial services (`<ANA>`). It then correctly identifies and recalls specific, relevant memories from the user's history, such as past complaints about high ATM fees and account limitations (`<MEM>`). By connecting the specific memory of "high fees" to the current option about "inability to withdraw money," the model makes a logical and accurate prediction. This demonstrates that OM-CoT successfully instills a more robust and transparent reasoning process, mitigating the issue of memory granularity mismatch.

Role Info:

- **Username:** celebrities
- **Description:** Welcome to your 15 seconds of fame! Just a bit of fun :)

Behavior History:

- **Post:**
  Action Time: 2020-08-06 13:13:54
  Content: Election 2020 #PresidentialDebates #PresidentTrump #Biden #USA
  Link: https://t.co/2SqNnemss9
- **Like:**
  Action Time: 2020-08-07 13:57:05
  Object: The Turning Point in @Cmdr_Hadfield's Life! Never give up and reach for the skies! #Space #Canada @csa_asc @esa @NASA
  Link: https://t.co/miadGaw46M
- **Like:**
  Action Time: 2020-08-14 09:59:57
  Object: He is without question a leader who pushes risky ideas forward. Via:@jongertner
  Link: https://t.co/ilyXah4F8n

Table 7: An example of FineRob data record

An example of GPT-4o augmented and reorganized data on Twitter

The user "mentoring" has shown a strong interest in mentoring-related content, particularly tweets that reference the mentoring movement led by @dianakimball. The user's behavior history indicates a pattern of liking and retweeting posts that express inspiration from the mentoring movement or that directly engage with it.

<ANA>

- Option (A) discusses thanking those who helped in a personal journey, which may resonate with the user's interest in mentorship but does not explicitly mention mentoring or the movement.
- Option (B) is a generic thank you for following, which is unlikely to engage the user as it lacks relevance to their interests.
- Option (C) explicitly mentions joining a movement and includes hashtags related to mentoring, making it highly relevant to the user's interests.
- Option (D) also mentions joining a movement but lacks the specificity of the mentoring context that option (C) provides.

Given the user's history of interactions, options (C) and (D) are the most relevant, but (C) stands out due to its explicit mention of inspiration and the mentoring movement.

</ANA>

<MEM>

The user's history includes multiple interactions with tweets that reference @dianakimball's mentoring movement. For example:

- The user liked and retweeted tweets that express inspiration from @dianakimball's movement, indicating a strong alignment with content that promotes mentorship.
- The user has engaged with tweets that directly mention the mentoring movement, such as "so inspired by @dianakimball's @mentoring movement" and "Want me to mentor you? I'm joining @dianakimball's @mentoring movement."

These interactions suggest that the user is actively looking for content that aligns with their interests in mentoring and is likely to engage with tweets that reflect similar themes.

</MEM>

Therefore, the answer is (C): "Inspired by, I'm now joining the movement. #Mentoring #ETD."

Table 8: An example of Reorganize Training data

Reasoning Case created by OM-COT-Finetuned Mistral-7b-Instruct on Reddit dataset

<ANA>

The options presented reflect varying perspectives on the accessibility and limitations of financial services, particularly regarding ATM withdrawals and online transactions. Option (C) specifically mentions the inability to withdraw money using a debit card, which aligns with the user's concern about the limitations of their financial account. Option (D) discusses the limitations set by financial institutions regarding ATM withdrawals and daily debt purchases, which is relevant to the user's inquiry about their financial situation.

</ANA>

<MEM>

- The user commented on the high fees associated with withdrawing money from an ATM, indicating a concern about the costs of accessing their funds.
- The user also expressed frustration over the limitations of their financial account, particularly regarding the inability to withdraw money without incurring fees.
- The user has engaged in discussions about the challenges of managing finances in Australia, which may relate to the limitations mentioned in option (D) regarding financial institution policies.

</MEM>

Therefore, the answer is (C).

Table 9: A reasoning case of OM-COT output.