Fine-Grained Behavior Simulation with Role-Playing Large Language Model on Social Media

Anonymous ACL submission

Abstract

Large language models (LLMs) have demonstrated impressive capabilities in role-playing tasks. However, there is limited research on whether LLMs can accurately simulate user behavior in real-world scenarios, such as social media. This requires models to effectively analyze a user's history and simulate their role. In this paper, we introduce **FineRob**, a novel fine-grained behavior simulation dataset. We collect the complete behavioral history of 1,866 different users on three social media platforms. Each behavior is decomposed into three fine-grained elements: object, type, and content, resulting in 78.6k QA records. Based on FineRob, we identify two dominant reasoning patterns in LLMs' behavior simulation processes and propose the **OM-CoT** fine-tuning method to enhance the capability. Through comprehensive experiments, we conducted an in-depth analysis of key behavior simulation factors and also demonstrated the effectiveness of the OM-CoT approach[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, accurately simulating human-like behaviors poses a substantial challenge, particularly in complex real-world scenarios.

In this work, we focus on simulating real social media user behavior. To explore how LLMs understand and simulate human behavior at a fine-grained level, we break down each behavior into three core elements: **object** (the target or recipient of the behavior), **type** (the nature of the behavior), and **content** (the specific details). This decomposition enables a detailed analysis of the model's ability to simulate various aspects of human behavior.

Specifically, we introduce the **FineRob** (**Fine**-Grained **Ro**le **B**ehavior) dataset, a novel benchmark for role-behavior simulation on social media. FineRob collects real user behavior data from three major platforms—Twitter, Reddit (primarily English), and Zhihu (primarily Chinese)—encompassing 1,866 distinct users and 78.6k fine-grained behavior elements. Each raw user behavior is decomposed into three sub-elements. 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.

Using the FineRob dataset, we conduct a comprehensive evaluation of nine widely used LLMs. Our analysis of the Chain of Thought (CoT) processes in behavior simulations revealed two prevalent reasoning patterns: "**role stereotype-based reasoning**" and "**observation and memory-based reasoning**".The former tends to over-rely on user profile information, which leads to reduced accuracy in behavior prediction. In contrast, the latter compares current observations with past behaviors, resulting in more accurate simulations, which is favored by more advanced models, such as GPT-4o.Building on this insight, we introduce a novel fine-tuning approach, OM-CoT, which utilizes special tokens to explicitly incorporate observation and memory analysis into the reasoning process. Extended experiments confirm the effectiveness of our approach.

---

[1] Code and dataset are available at `https://anonymous.4open.science/r/FineRob-791B/`

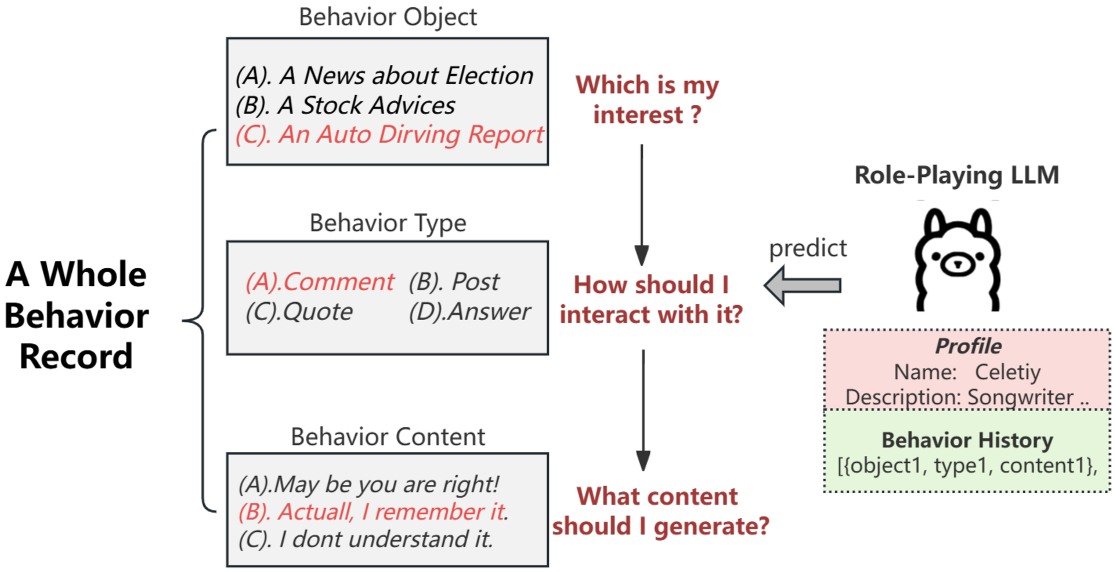

Behavior Object
(A). A News about Election
(B). A Stock Advices
(C). An Auto Dirving Report
Which is my interest ?
Behavior Type
(A).Comment (B). Post
(C).Quote (D).Answer
How should I interact with it?
Behavior Content
(A).May be you are right!
(B). Actuall, I remember it.
(C). I dont understand it.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

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.

The contributions of this study can be summarized as follows:

- **We introduce the FineRob dataset**, consisting of 78.6k fine-grained behavior element prediction QA records from 1866 real-world social media users. FineRob serves as a realistic, multilingual benchmark for evaluating LLMs' ability to simulate role-specific behaviors.
- **We comprehensively assess behavior simulation** across nine mainstream LLMs, focusing on the reasoning patterns employed during simulation.
- **We propose the OM-CoT, a novel fine-tuning method** that explicitly integrates observation and memory analysis into the reasoning process using special tokens. Our experiments demonstrate significant improvements across all three behavior element prediction tasks.

2 Related Work

2.1 Role-Playing LLM

Recently, numerous RP-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).Beyond this, techniques such as in-context learning (ICL)(Zhao et al., 2024) and retrieval-augmented generation (RAG)(Liu et al., 2023) have been employed. Additionally, supervised fine-tuning on targeted dialogue datasets(Shao et al., 2023; Zhou et al., 2023; Tu et al., 2024) and methods like LoRA(Yu et al., 2024) have further enhanced RP-LLMs' role-playing capabilities. Despite these advancements, research in RP-LLMs remains in its early stages, with a primary focus on mimicking conversations.

2.2 LLM for User Behavior Simulation

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 reinforce-

---

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

Dataset	Source	Size	Usage	Language	Type	isReal?	Fine-Grained
PersonalDialog(Zheng et al., 2019)	Weibo	20.8M	Train&Test	ZH	Dialogue	✓	–
Ditto(Lu et al., 2024)	LLM Synthetic	4k	Train&Test	EN,ZH	Dialogue	×	–
LaMP-7(Salemi et al., 2024)	Twitter	12k	Train&Test	EN	Dialogue	✓	–
PIPPA(Gosling et al., 2023)	Character.AI	26k	Train	EN	Dialogue	×	–
ROCStories(Mostafazadeh et al., 2016)	Little Stories	98K	Train&Test	EN	Behavior	×	×
choices13K(Bourgin et al., 2019)	gamble	13K	Test	EN	Behavior	✓	×
Life-Choice(Xu et al., 2024)	Novels	1.4K	Test	EN	Behavior	×	×
FineRob	SocialMedias	78.6K	Train&Test	EN,ZH	Behavior	✓	✓

Table 1: Comparison between FineRob and previous role-playing datasets.

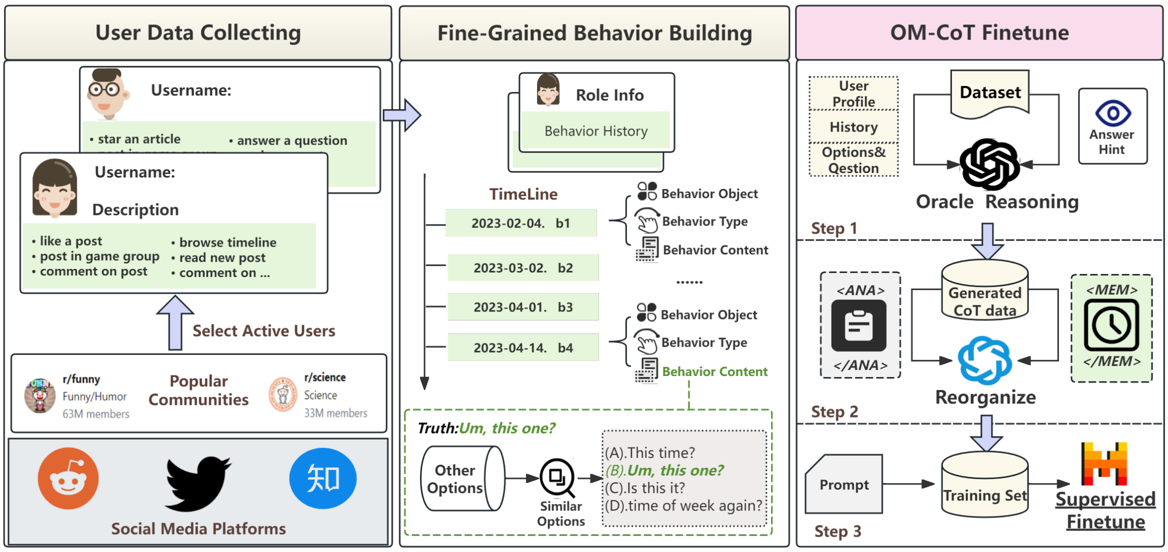


Figure 2: Overview of our work, The left and middle sections of the figure illustrate the process of constructing the FineRob dataset. The right section shows how OM-COT-FineTune training details, including data augmentation, reorganize with special tokens and SFT training.

ment learning algorithms to optimize recommendation systems based on user feedback simulated by LLMs. The work by (Xu et al., 2024) is particularly relevant, which introduced the "LIFECHOICE" 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.

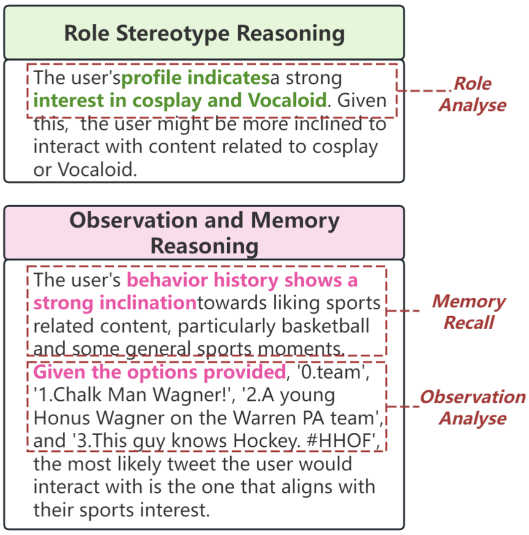
3 FineRob Dataset

3.1 Data Collecting

Our goal is to explore how LLMs simulate the behaviors of real internet users. To achieve this, we focused on social media platforms, including Twitter (now X), Reddit, and the Chinese question-and-answer website Zhihu. From these platforms, we collected extensive behavioral histories of real users, as shown in the first part of Figure 2. Unlike other role-playing tasks, we emphasize fine-grained behavior simulation in real-world scenarios. A detailed comparison is provided in Table1.

Principles Our data collection strategy is guided by several key principles. (1) **Popularity**: We focused on mainstream, widely discussed topics and scenarios to ensure the dataset reflects a representative sample of user behaviors.(2) **Diversity**: We include a broad range of user profiles and behavioral patterns to enhance the generalizability of our findings.(3)**Activity**: We select users who are active within the community and engage in various types of behavior, helping to minimize data contamination from social bots, fake accounts, or other non-human users.

User Selection To collect our dataset, we target active users by selecting them from trending topics or communities displayed on aggregation page of each platform. For Reddit, we choose popular posts from the top 20 communities and filter



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 3: 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.

participants based on predefined principles. Using the PRAW[3], we automatically collect these users' timelines, including their complete historical behaviors.For Twitter, we purchase access to the official API[4] and prioritize users who actively engaged in discussions on trending topics. On Zhihu, we focus on users who frequently ask or answer questions. To ensure reliability, we filter users with at least 70 accessible historical behaviors. Conversely, we exclude users who exhibit excessive activity, as these accounts may be operated by multiple individuals.

3.2 Fine-Grained Behavior Building

Next, we convert the raw user timelines into a fine-grained behavior prediction QA dataset with multiple-choice format. Specifically, each behavior record is broken down into three elements: **object** (the recipient of the behavior), **type** (the nature of the behavior), and **content** (the specific details of the behavior). This process is illustrated in the middle section of Figure 2. 2.

A significant challenge lies in constructing valid alternative options for each behavior element in the multiple-choice format. The behavior **type** options are relatively straightforward, as platforms typically have predefined actions like "Post" and "Comment." These behavior types can be found in Appendix 6.For the behavior's object and content options, we construct a candidate set based on the user's active times and communities. From this set, we calculate the similarity to the correct option using sentence embeddings[5]. To increase task difficulty and introduce ambiguity, we randomly sample three options with closely aligned sentiment to the correct answer.A detailed entry can be found in Appendix 7.

Finally, we collected a total of 78.6k behavior element prediction records and split the dataset into approximately 61k for the training set and 17.6k for the test set. To assess the generalization capability of large language models (LLMs), we ensured that no user roles overlap between the training and test sets.

4 Methodology

4.1 Preliminary Analysis

We conduct preliminary experiments using a zero-shot Chain of Thought (CoT) approach.Our goal is to understand the reasoning processes LLMs use in behavior simulation tasks.Our analysis reveal two primary reasoning patterns. The first, termed role stereotype-based reasoning, derives outcomes by analyzing character profiles.The second, **observation and memory-based reasoning**, involves analysing all observed options and linking them to similar past scenarios.Examples of these patterns are illustrated in Figure 3.

In our comparison, we find that more advanced models, such as GPT-4o, tend to prefer the "observation and memory-based reasoning" pattern, leading to more accurate simulations. To investigate this further, we conduct a quantitative analysis of the similarity between CoT reasoning text and each parts of prompt: behavior history, observed options, and role info. The results, shown in Figure 4, reveal an interesting insight: A higher similarity to role profiles, which may involve more character analysis, does not always lead to better behavior simulations. Instead, focusing on historical data and observed options proves to be more effective in improving simulation accuracy.

4.2 OM-CoT Finetune

Based on these findings, we propose a straightforward yet effective method called OM-CoT Finetune

---

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

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

[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.

(Observation and Memory-based Chain of Thought Finetune) to enhance behavior simulation accuracy in LLMs.This method promotes the "observation and memory-based reasoning" pattern by explicitly integrating observation and memory analysis into the CoT reasoning. The approach involves three steps, as illustrated in the right part of Figure 2.

Oracle CoT Generation First, we use a powerful large language model to generate CoT reasoning. To prevent error propagation, we adopt an oracle setting, where the correct answer is provided in the input prompt. This ensures the model references the correct behavior during reasoning. We carefully adjust the prompt to make sure the generated CoT does not inadvertently reveal the correct answer.

Reorganize CoT with special tokens Next, we introduce two special tokens: `<ANA>` and `<MEM>`. A smaller LLM reorganizes the CoT results by wrapping observation-based analysis within `<ANA></ANA>` and historical behavior analysis within `<MEM></MEM>`. At the end of each reasoning process, the model explicitly states the final behavior decision (e.g., "Therefore, the behavior type is A.Comment"). A detailed case can be found in Appendix 8

SFT with Enhanced Dataset Finally, we perform Supervised Fine-Tuning (SFT) on the LLM using the reorganized dataset. The training optimizes for standard language model loss, guiding the model to effectively utilize the special tokens along with system prompts to control the CoT process. We create 60K instruction training data for OM-CoT fine-tuning, which is also available in the codebase.

5 Experiment

5.1 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: zero-shot, few-shot, standard-CoT fine-tune, and OM-CoT fine-tune. In the few-shot setup, we include a reasoning example created by GPT-4o, which follows the "observation and memory-based reasoning" pattern. 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

---

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

	Reddit			Twitter			Zhihu		
	BO	BC	BT	BO	BC	BT	BO	BC	BT
Mistral-7b-Instruct									
Zero-Shot	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
Few-Shot	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
std-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									
Zero-Shot	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
Few-Shot	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
std-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.

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.

5.2 Main Result

We first compare the behavior simulation capabilities of the main LLMs under the same zero-shot setting, as shown in Table2. Following this, we perform extended experiments to highlight the advantages of the OM-CoT fine-tuning method, with the results presented in Table 3. Next, we will discuss some conclusions drawn from the main results.

Commercial Closed-Source Models Still Perform Better While many open-source models have recently demonstrated strong performance on general leaderboards[9], a detailed comparison in Table 2 shows that even the best open-source LLMs lag behind the GPT-4 series by approximately 15% in average F1 score. Comparable performance is only observed on the Reddit dataset for Behavior Object and Behavior Content prediction tasks. This suggests that role-playing and behavior simulation may require more than just the reasoning abilities. These tasks likely involve advanced capabilities such as empathy and reflective analysis of past behaviors.

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 As shown in Table 3, we apply OM-CoT fine-tuning to two models with different parameter sizes, a reasoning case can be found in Appendix 9. For the Mistral-7B model, performance improve across all nine sub-tasks, with an average F1 score increase of approximately 4.5%. Similarly, the Solar-10.7B model exhibited gains in seven out of nine tasks, including a significant 9.8% improvement in the Reddit Behavior Object prediction task. Furthermore, baseline analysis reveal that incorporating examples from the fine-tuning

---

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

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

[9] `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`

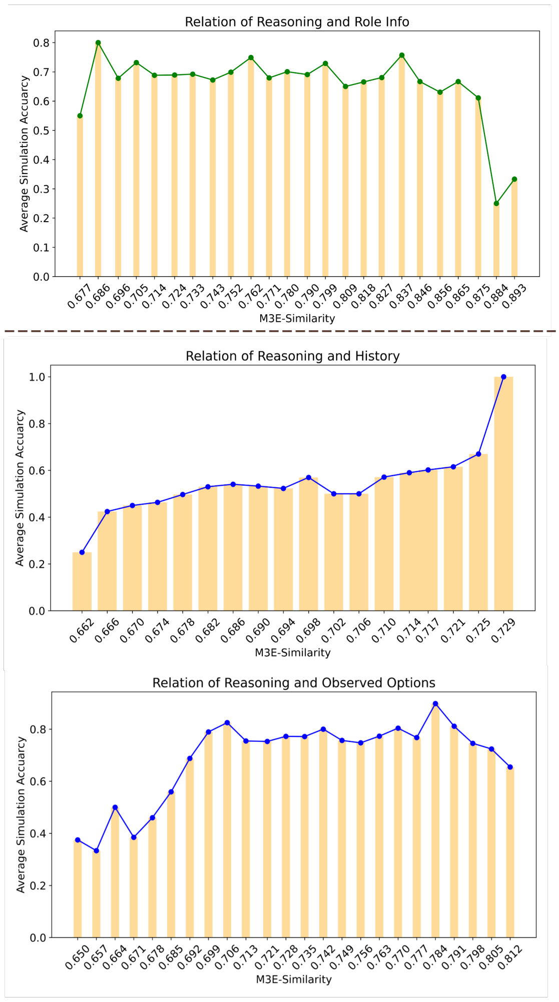


Figure 4: Analysis of simulation accuracy changes across different similarity levels between reasoning and various parts of the prompt. The results are generated using ChatGPT-3.5-turbo-0125 on the Twitter test set, with the average F1-score calculated across three behavior element tasks.

dataset into the few-shot setting do not produce the expected improvements; in some cases, it even perform worse than the zero-shot setting. On the other hand, both Std-CoT-FT and OM-CoT-FT consistently outperform models that were not fine-tuned. These results indicate that, even with different users in the training and test sets, large models can effectively learn generalizable reasoning patterns.

LLMs Struggle with Short-Behavior Tasks, Even When Fine-Tuned 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: Does adding more user history input improve the accuracy of behavior simulation? In the main experiment, we consistently choose the 30 most recent behavior history as the input. However, a plausible hypothesis suggests that includ-

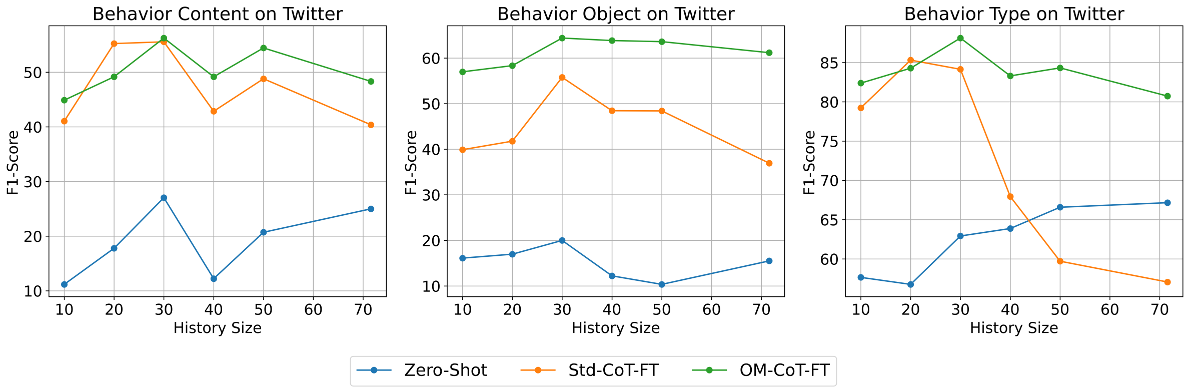


Figure 5: The relationship between input historical behavior size and the accuracy of simulating fine-grained behavior elements. The figure presents the results of three methods on the Twitter dataset.

ing more behavior history could enhance behavior simulation, provided it fits within the model's token limit.To explore this, we evaluate the performance across different history window sizes, ranging from 10 to all entries(average 74), as shown in Figure 5.Contrary to intuition, **adding more user behavior history does NOT consistently improve behavior simulation.** We find that performance peaks at around 30 behavior entries, with additional history leading to a decline in accuracy. We hypothesize that increasing historical data introduces more noise, making it harder for the model to focus on relevant information. Interestingly, OM-CoT-FT models display greater stability with increased input history, showing promise for handling longer behavior sequences.

	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.

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.

6 Conclusion

In this work, we introduce FineRob, a new fine-grained behavior simulation dataset designed to explore how LLMs understand and simulate real human behaviors. We collect user behavior history from three social media platforms and break down each behavior record into three key elements: object, type, and content. Preliminary experiment on FineRob reveals two dominant reasoning patterns: role stereotype-based, observation and memory-based. The latter proves to be more effective in behavior simulation, leading us to propose OM-CoT method that explicitly integrates observation and memory into the reasoning process.We conduct comprehensive experiments on nine mainstream LLMs and analyze key influencing factors in depth. By comparing OM-CoT with baseline methods, we confirm the effectiveness of our approach.

Limitations

A limitation of the FineRob dataset lies in the handling of building behavior object task's options. Although we reduce the candidate set through user's interest communities and active time ranges, it remains difficult to determine whether a given incorrect option is "observing but not interacting" or "not observing at all." This ambiguity arises because browsing history is private and inaccessible. Additionally, we did not include a "None of the above" option, which could be valuable in exploring the boundaries of character behavior. We plan to address this in future research.

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.

Shuai Liu, Hyundong Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. 2023. RECAP: retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. In *ACL (1)*, pages 8404–8419. Association for Computational Linguistics.

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.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *EMNLP*, pages 13153–13187. Association for Computational Linguistics.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *CoRR*, abs/2401.01275.

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*.

Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. Neeko: Leveraging dynamic lora for efficient multi-character role-playing agent. *CoRR*, abs/2402.13717.

Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. 2024. Narrative-play: Interactive narrative understanding. In *EACL (Demonstrations)*, pages 82–93. Association for Computational Linguistics.

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

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

A behavior content QA record of Twitter

Role Info:

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

Interests:

- Swachhsurvekshan
- Ogwugfood
- Foodapp
- Survey
- EcoVizag
- Fooddelivery

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
- **Retweet:**
  Action Time: 2020-08-14 10:02:52
  Object: RT @savanteum: He is without question a leader who pushes risky ideas forward.
- **Post:**
  Action Time: 2020-09-16 11:24:00
  Content: Oh Trump...
  Link: https://t.co/uYdrgSORfn
- **Post:**
  Action Time: 2020-11-16 22:51:14
  Content: https://t.co/NbusQ7o6DT
- **Post:**
  Action Time: 2020-11-20 21:00:28
  Content: :)

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: An example of OM-COT Finetuned Reasoning COT