PoliSim: Evaluating Large Language Model-based Agents in Politician Simulation

Anonymous ACL submission

Abstract

The capabilities of Large Language Models (LLMs) to model and imitate humans offer new perspectives for simulating politicians and society. However, the specific aspects and extent to which LLMs can effectively simulate politicians remain unexplored. Previous evaluations have primarily focused on fictional characters and superficial characteristics, such as linguistic styles, while ignoring LLMs' capacity to accurately replicate individuals' complex features, such as their opinions and actions. This paper introduces PoliSim, a novel benchmark designed to comprehensively and objectively assess the effectiveness of politician simulation by LLM-driven agents. Grounded in cognitive behavior theory, PoliSim evaluates simulations across cognition, attitude, and behavior. By utilizing data from 1,000 politicians, PoliSim transforms the information into a unified evaluation framework consisting of multiple-choice and generation questions. We apply PoliSim to various LLMs and simulation schemas to offer insights and directions for future research in realistic agent-based simulations.

1 Introduction

Research on political actor modeling focuses on developing computational techniques to analyze and predict the language and behavior of political figures (Kornilova et al., 2018; Feng et al., 2022). Among these actors, politicians are particularly crucial, as their decisions and communications directly shape governance and public opinion. Traditional statistical or deep models, trained on various data (Budhwar et al., 2018; Yang et al., 2021; Mou et al., 2023), are limited to classifying individuals with discrete labels. In contrast, recent advancements in Large Language Models (LLMs) (Xi et al.; Wang et al., 2024) enable agents to simulate more complex human behaviors, such as mimicking responses to specific situations, revolutionizing the modeling of politicians.

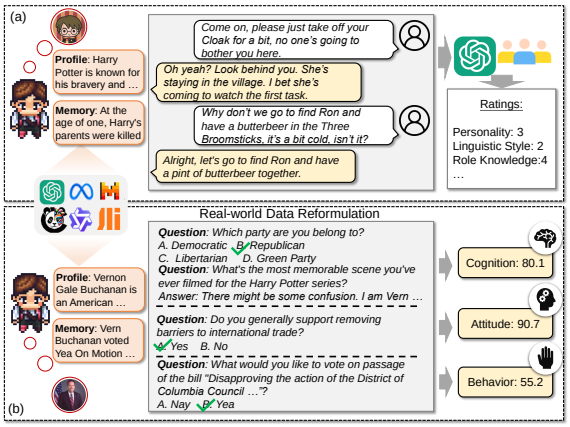

Figure 1: (a) An illustration of previous virtual character role-playing evaluation, which relies on ratings given by GPT or humans. (b) An illustration of our proposed PoliSim. We reconstruct real-world data to evaluate the performance of agents across various dimensions.

With the potential to imitate humans, LLMs have been utilized to empower social simulation (Park et al., 2022, 2023, 2024), which can be applied to various scenarios, such as policy making (Xiao et al., 2023), election poll (Argyle et al., 2023) and public opinion mining (Törnberg et al., 2023; Mou et al., 2024c). In these scenarios, whether LLM-based agents can authentically simulate individuals is at the core of the replication of the actual situations. Although some works have evaluated LLMs' role-playing capabilities (Chen et al., 2024b) to simulate specific characters, these attempts are mostly limited to (1) **virtual scenarios** such as novels and scripts (Chen et al., 2023; Xu et al., 2024) and (2) **surface-level features** such as linguistic styles (Wang et al., 2023b) and role-specific knowledge (Shao et al., 2023), as shown in Figure 1(a) and Table 1. This leaves a gap in whether LLMs can replicate more complex attributes of politicians, such as opinions and well-considered actions, hindering the simulations for real-world applications.

To bridge these gaps, we introduce **PoliSim**, a

Work	chars	# chars	Eval dims	Method
CharacterLLM (Shao et al., 2023)	real	9	memorization, values, personality, hallucination, stability	Subjective
RoleLLM (Wang et al., 2023b)	virtual	100	speaking style, response accuracy, role knowledge	Subjective
CharacterGLM (Zhou et al., 2023)	virtual	250	consistency, human-likeness, engagement	Subjective
LifeChoice (Xu et al., 2024)	virtual	1,401	decision-making	Objective
PoliSim (ours)	real	1,000	cognition, attitude, behavior	Objective

Table 1: Comparison between PoliSim and previous works.

benchmark designed to comprehensively and objectively assess the effectiveness of LLM-based agents for politician simulation in real scenarios. The foundation of the evaluation is built on two key questions: (1) **What** dimensions should be considered to observe the agents simulating politicians? and (2) **How** to evaluate the fidelity, i.e., the alignment with the corresponding individuals of different simulation methods? For the first question, we adopt the cognitive behavior theory (Beck, 1979) to identify three interrelated dimensions for analyzing humans: *cognition*, *attitude* and *behavior*. For the second question, to avoid the instability and high costs of subjective methods like GPT-4 or human ratings, we construct questions linked to these dimensions for a wide range of politicians. However, due to the sparsity of real-world data, it is difficult to directly collect evaluation questions. For cognition, we design multiple-choice questions and specific generation questions based on personal experiences to evaluate whether agents reflect the cognitive boundaries of their counterparts. For attitude, we reconstruct questionnaires and statements to test whether the agent's ideology and opinions align with those of the individual. For behavior, we re-formulate the legislative data to see whether the agents express group preferences and take actions consistent with the corresponding individuals. Ultimately, we have developed 37,887 questions covering 1,000 politicians.

We conduct a comprehensive evaluation on various LLMs and simulation schemas. Our analysis reveals that activating a robust base model, such as GPT-4 (OpenAI, 2023) or Llama-3 (Team, 2024), along with profiles and memory, produces the best results across multiple dimensions. For smaller models, incorporating memory proves to be particularly beneficial. However, models generally struggle with recognizing and managing their cognitive boundaries. The most significant challenge across all models is behavior replication, although improvements in attitude modeling can enhance behavior simulation. Additionally, we examine factors influencing simulation effectiveness, including LLMs' internal reasoning processes and the methods of knowledge provision.

In summary, our contributions include:

- We first explore politician simulation using LLMs and employ the cognition behavior theory to observe agents and provide a comprehensive evaluation of political simulation.
- We construct a benchmark for politician simulation based on real scenarios, paving the way for evaluating and improving the abilities of LLMs to simulate humans.
- We conduct thorough experiments and analysis on politician simulation, providing insights for the optimization of the simulation schema.

2 Related Work

2.1 Political Actor Modeling

Political actor modeling focuses on modeling the attributes and behaviors of political actors. Previous work mainly applies statistical methods (Clinton et al., 2004; Gu et al., 2014; Vafa et al., 2020) or deep models (Kornilova et al., 2018; Mou et al., 2021; Feng et al., 2022; Mou et al., 2023) to estimate the ideology or votes of political actors. Despite improvements in target tasks, they are limited to the training and testing schema. Benefiting from the strong generalization capabilities, LLMs have demonstrated the potential to model political actors in zero-shot settings (Wu et al., 2023; Mou et al., 2024b). However, the full extent and nature of LLMs' ability to simulate political actors remain largely unexplored.

2.2 LLM-based Social Simulation

LLMs have recently been applied to simulate social dynamics, aiming to replicate real-world phenomena (Mou et al., 2024a). Efforts have been made to simulate users in recommendation systems (Wang et al., 2023a), opinion dynamics (Yang et al., 2024) and fake news propagation (Liu et al., 2024). These

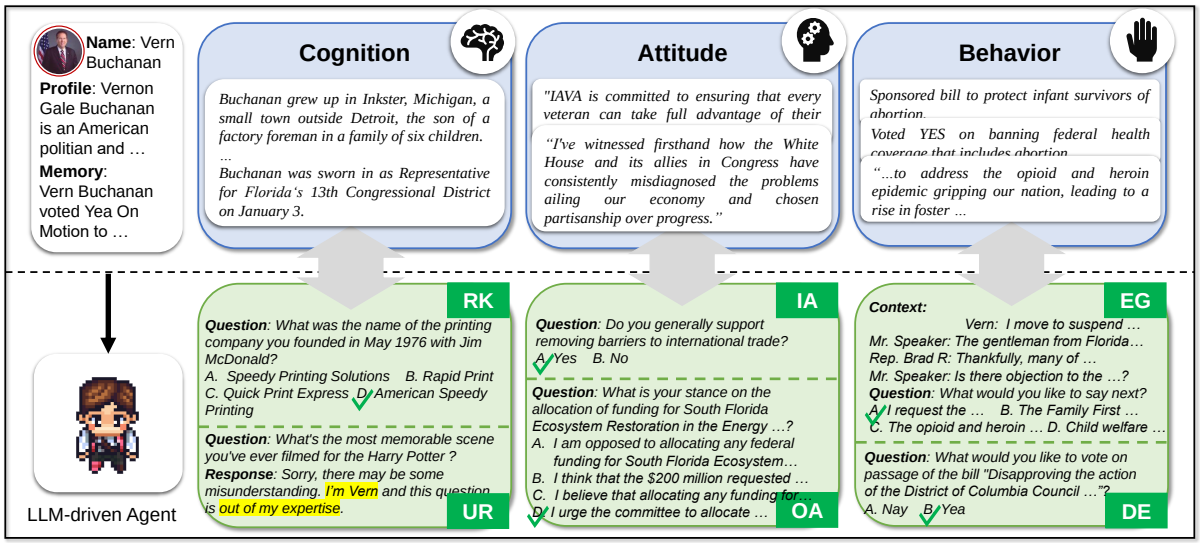

Figure 2: An illustration of PoliSim. PoliSim consists of 6 types of questions re-formulated from real-world data to evaluate the simulation of political actors in cognition, attitude and behavior.

simulations often focus on how LLMs can replicate human-like responses and attitudes. Although LLMs can generate realistic and contextually relevant outputs, challenges still remain, such as ensuring the accuracy of simulated behaviors and addressing the potential biases and hallucinations inherent in LLMs (Guo et al., 2024).

2.3 Evaluating Role-playing Agents

Evaluation of LLM-based role-playing agents can be divided into two categories. One line of research focuses on assessing role-playing capabilities that are not tied to specific personas, such as engagement (Zhou et al., 2023), emotion understanding (Huang et al., 2023a) and problem-solving ability (Xu et al., 2023). Another line concentrates on persona fidelity, i.e., whether the agents can replicate the intended personas. They mainly focus on the alignment with regards to role knowledge (Shao et al., 2023), linguistic style (Wang et al., 2023b) and personality(Huang et al., 2023b), but ignore more complex characteristics such as behaviors. Although very recently Xu et al. test decision-making, they focus on characters in novels and thus may simplify the challenging situation of simulation of real human choice.

3 PoliSim: Benchmark for Politician Simulation

We present the PoliSim benchmark in this section. First, guided by cognitive behavior theory, we identify the dimensions to be observed. Then, we construct multi-dimensional evaluation questions based on real-world data. Different from previous work, our focus shifts from evaluating conversational abilities as in dialogue systems to assessing performance across critical dimensions of responses and interactions in various scenarios.

3.1 Observe Agents using Cognition Behavior Theory

To evaluate how closely an agent resembles the person it simulates, we first need to determine the key aspects that define such resemblance. Previous work has mostly regarded agents as dialogue systems (Zhou et al., 2023) rather than substitutes for humans in social research (Park et al., 2024), focusing on isolated dimensions such as speaking style and role-specific knowledge. However, the aspects emphasized when observing human behavior in social experiments and surveys differ significantly. The cognitive behavior theory (Beck, 1979; Adler, 2014) provides a useful framework by exploring internal mental processes and conceptualizing human psychological functions as interconnected systems of thoughts, emotions, and behaviors. Building on this theory, we identify three dimensions, i.e., cognition, attitude, and behavior.

3.1.1 Cognition Dimension (Cog.)

In the cognition dimension, agents are expected to exhibit cognitive boundaries similar to those of the individuals they are simulating. This should be measured in two aspects:

Role Knowledge (RK). Agents should demon-

strate an accurate awareness of their own role and the surrounding world. We evaluate this by testing whether agents can answer role-specific questions correctly, without factual errors or hallucinations.

Unknown Knowledge Rejection (UR). Agents should recognize and refuse to answer questions that fall outside their cognitive boundaries due to factors like age, era, or personal experience (Lu et al., 2024). This tests the agents' ability to reject questions that are beyond their knowledge, thereby minimizing hallucinations.

3.1.2 Attitude Dimension (Att.)

In addition to cognition, we expect the agents to have emotional attitudes consistent with the corresponding individuals. This alignment is crucial for applications in social simulations such as public opinion analysis and legislative processes (Baker and Azher, 2024). We measure attitude alignment through both coarse and fine-grained aspects:

Ideology Alignment (IA). Ideology grasps the overall political leaning or preferences issues such as abortion and immigration (Liu et al., 2022; Mou et al., 2023). Agents are expected to display consistent ideological positions, such as conservative or liberal, and to support or oppose issues in line with the individuals they simulate.

Opinion Alignment (OA). Compared to the discrete position provided in ideology, fine-grained opinions reflect more specific thoughts of individuals. Opinion alignment assesses whether agents maintain consistency with the designated political actors when responding to specific questions.

3.1.3 Behavior Dimension (Beh.)

Behavior is the ultimate external manifestation of an individual's traits. Whether agents could take actions consistent with the intended personas is challenging but indispensable for real-world applications. However, it remains an under-explored question. To fill this gap, we propose to observe the performance at the behavior dimension from expression in group dynamics and decision-making.

Expression in Group Dynamics (EG). Individuals can exhibit complex behaviors in group conversation scenarios, such as controlling the pace of the discussion and considering others' perspectives (Chen et al., 2024a). Agents need to maintain consistent preferences and responses when engaged in sophisticated group conversations.

Decision-making (DE). When being at the decision points, well-established agents should make the same choice as the corresponding individuals. This requires agents to accurately identify and reason about the individuals' attitudes and behaviors relevant to the scenario.

3.2 Construction of PoliSim

In this part, we introduce the construction of PoliSim. As illustrated in Figure 2, we build our dataset on real data from various sources. Then we re-formulate the data into questions to cover the dimensions outlined before.

3.2.1 Data Collection

Our data collection consists of two main components: profile and memory of politicians, and scenario data for evaluation.

Profile and Memory The profile is the description of the corresponding individual. For politicians, we collect their biography from the website of Congress [1] and follow Wang et al. to summarize the profiles into natural languages using GPT-3.5. For memory, since real people do not have a complete storyline like fictional characters in novels, we pay attention to records of words and deeds, by collecting their historical statements and voting records before 2023 from Twitter (Mou et al., 2023), VoteSmart, and Legiscan [2]. More details can be found in Appendix A.

Evaluation Data Source Following (Shao et al., 2023; Lu et al., 2024), we collect wiki pages of politicians to construct the cognition-related questions. For the attitude dimension, we collect political courage tests [3] for issue positions and statements after 2023 from VoteSmart. For the behavior dimension, we collect roll-call voting records after 2023 from Legiscan and parsed congressional records (Gentzkow et al., 2018) of the 112th to 114th sessions to get the legislative debate records.

3.2.2 Formulation of the Questions

We reformulate the raw data into multiple-choice questions and specific generation questions to enable automatic and objective model evaluation.

Cognition Dimension To evaluate cognition, we adopt a reading comprehension approach (Lu et al., 2024), consisting of three key steps. (1) Question Generation: We provide GPT-3.5 with wiki pages of both the target individual A and a comparison individual B. GPT-3.5 generates two

[1] https://www.congress.gov/
[2] https://legiscan.com/
[3] https://justfacts.votesmart.org/

types of questions: positive questions that A can answer to test A's RK dimension and negative questions that B can answer but A cannot, to test A's cognitive boundaries, i.e., the UR dimension. (2) Answer Generation: For RK, the positive questions are appended after A's wiki description, and GPT-3.5 extracts relevant information to emulate A's responses, which serve as the correct answers. (3) Option Generation: For each positive question, GPT-3.5 generates three additional distractor options to form multiple-choice questions. Detailed are included in Appendix B.

Attitude Dimension To test the attitude dimension, we use political courage test data with predefined reference answers to evaluate IA. For OA, to address the misalignment between model-generated broad questions and specific statements as answers, we use a three-step process: (1) extract the core opinion from statements, (2) generate specific questions based on the opinion, and (3) create three distractor options by rephrasing the opinion with alternative arguments or stances.

Behavior Dimension For EG, we used legislative debate data, providing scenario descriptions and multi-turn discussions as context. The original responses serve as correct answers, while GPT-3.5 rephrased responses with different preferences as distractors. For DE, we used post-2023 roll-call voting data, asking agents how they would vote on each bill. Actual voting results are correct answers, with potential voting options as candidates.

3.2.3 Data Validation and Analysis

Manual Filtering Since we construct the data with the help of LLMs, it is necessary to manually review the data to remove any unreasonable entries. For UR questions, we manually remove those questions that the individuals could answer or for which we were unsure whether they could answer, such as "*What's the most challenging role you've ever taken on?*". For RK, OA, and EG questions, we removed negative options that had noticeable differences in length or discourse markers compared to the correct answers, to avoid the model making inferences based on these shortcuts.

Data Analysis We show the statistics of PoliSim in Table 2. Except for the EG questions, which include multi-turn dialogue history in the debate, all other evaluations can be conducted through a single round of Q&A. Due to data sparsity, only 101 politicians have complete evaluation data across all dimensions, but this subset, named PoliSim-role, is still comparable to previous works (Wang et al., 2023b; Chen et al., 2024a) in size. More details about data can be found in Appendix A.

	Cognition		Attitude		Behavior	
	RK	UR	IA	OA	EG	DE
Metrics	Acc.	Rej.	Acc.	Acc.	Acc.	Acc.
# Characters	917	937	912	253	500	381
# Questions	4,226	4,630	15,822	4,739	4,328	4,142
# of options per Question	4	N/A	2	4	4	2
Avg. Instruction Len. (words)	14.34	14.56	12.69	16.48	10.48	40.54

Table 2: Metrics and statistics of PoliSim. There are 1,000 political actors in total, including 37,887 questions. Among them, 101 individuals have evaluation questions in all the dimensions, including 7,725 questions. We call this subset PoliSim-role.

4 Experiment Settings

4.1 Evaluation Metrics

Previous works (Shao et al., 2023; Wang et al., 2023b) often rely on LLMs or human evaluations, leading to unstable results and high costs. In PoliSim, all tasks except for UR are constructed as multiple-choice questions. Following the MMLU(Hendrycks et al., 2020), we use **accuracy (Acc.)** as the evaluation metric. For UR, where models are required to generate complete answers, we calculate the **rejection rate (Rej.)**. We use rule-based methods and an LLM judge to determine whether the model correctly rejects the question. Notably, if a model refuses to answer by stating it is an AI model, this will not be counted as a valid refusal since it deviates from the persona.

4.2 Reasoning Models

We use the following model as the base of the agent for testing: **Llama-2-7B-chat-hf** (Touvron et al., 2023), **Llama-3-8B-Instruct** (Team, 2024), **Vicuna-7B-v1.5** (Zheng et al., 2023), **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023),**ChatGLM2-6B** (GLM et al., 2024), **GPT-3.5-Turbo** (OpenAI, 2022), and **GPT-4-Turbo** (OpenAI, 2023). We focus on open-source models around 7B because most existing role-playing models (Shao et al., 2023; Yu et al., 2024) are of this size.

4.3 Simulation Methods

We include three mainstream methods for constructing agents to simulate individuals.

Profile Using the profile of the individuals as the system prompt has been widely adopted by previ-

Model	Sample-Level							Role-Level						
	Cognition		Attitude		Behavior		Avg.	Cognition		Attitude		Behavior		Avg.
	RK	UR	IA	OA	EG	DE		RK	UR	IA	OA	EG	DE	
	Acc.	Rej.	Acc.	Acc.	Acc.	Acc.		Acc.	Rej.	Acc.	Acc.	Acc.	Acc.	
Only Profile														
Llama-2-7B-chat-hf	61.8	38.9	62.8	83.7	45.8	53.8	57.8	57.1	46.9	69.7	84.9	47.5	39.5	57.6
Llama-3-8B-Instruct	76.0	59.6	82.0	92.1	80.5	62.3	75.4	73.8	77.8	85.2	**94.8**	81.3	50.4	77.2
Vicuna-7B-v1.5	62.4	13.6	66.3	84.0	71.0	54.4	58.6	60.6	13.4	74.6	84.5	73.5	43.0	58.2
Mistral-7B-Instruct-v0.3	70.2	58.3	77.3	90.0	71.2	58.8	71.0	70.4	83.2	83.0	89.9	72.4	46.5	74.2
ChatGLM2-6B	45.9	9.7	58.9	74.8	45.9	50.6	47.6	43.7	9.2	67.9	76.9	55.1	49.3	50.3
GPT-3.5-Turbo	70.8	67.6	70.7	90.9	80.5	67.7	74.7	73.0	89.0	86.3	90.5	85.4	58.9	80.5
GPT-4-Turbo	**85.6**	75.8	85.6	**94.3**	**84.4**	**77.2**	**83.8**	78.8	80.4	91.2	93.8	**87.9**	75.0	84.5
Avg.	67.5	46.2	71.9	87.1	68.5	60.7	67.0	65.3	57.1	79.7	87.9	71.9	51.8	68.9
Profile+Memory														
Llama-2-7B-chat-hf	61.3	49.2	65.3	81.4	54.5	57.4	61.5	60.8	48.1	68.6	80.7	59.4	49.8	61.2
Llama-3-8B-Instruct	71.2	85.4	85.0	93.0	82.2	65.9	80.4	**82.3**	87.2	84.0	92.5	85.0	68.9	83.3
Vicuna-7B-v1.5	68.5	54.5	75.9	87.8	72.0	60.4	69.9	63.0	50.9	78.6	88.0	74.1	59.1	68.9
Mistral-7B-Instruct-v0.3	69.0	70.5	81.2	90.1	72.7	63.5	74.5	67.6	74.9	81.7	89.5	78.5	59.3	75.3
ChatGLM2-6B	52.6	28.7	59.2	81.0	46.0	51.0	53.1	54.7	27.1	64.0	81.1	55.1	48.2	55.0
GPT-3.5-Turbo	77.2	80.7	84.3	92.2	78.7	63.2	79.4	77.0	73.3	90.8	94.1	86.1	**80.3**	83.6
GPT-4-Turbo	84.3	73.6	88.6	93.8	82.1	74.6	82.8	77.4	85.4	**91.4**	93.0	86.9	77.1	**85.2**
Avg.	69.1	63.2	77.1	88.5	69.7	62.3	71.7	69.0	63.8	79.9	88.4	75.0	63.2	73.2
Role-playing Models														
CharacterGLM-6B	45.3	10.9	59.1	74.4	50.1	51.1	48.5	43.2	10.2	68.5	75.0	57.5	46.6	50.2
Ditto-Llama3	78.3	**94.1**	79.0	93.9	78.5	57.5	80.2	74.2	**97.0**	81.9	93.5	77.0	45.5	78.2
Neeko-Llama3	59.8	14.6	70.3	28.6	19.6	54.4	41.2	37.6	14.8	29.0	44.5	12.0	39.9	29.6
Baichuan-NPC-Turbo	73.6	85.9	79.4	89.4	68.3	52.8	74.9	71.1	84.8	82.8	90.1	81.7	46.3	76.1
Xingchen-plus	77.3	66.4	**88.7**	92.1	76.0	76.5	79.5	81.4	65.1	86.8	90.1	67.7	66.0	76.2
Avg.	66.9	54.4	75.3	75.7	58.5	58.5	64.9	61.5	54.4	69.8	78.6	59.2	48.8	62.0

Table 3: Main results of PoliSim evaluation. *Sample-level* columns present the results of evaluation on samples of PoliSim, while *Role-level* columns report the averaged results of individuals in PoliSim-role. Best performances are shown in *bold*, and suboptimal ones are *underlined*.

ous works (Wang et al., 2023b; Lu et al., 2024). This approach primarily mimics individuals leveraging the internal knowledge of LLMs.

Profile with Memory Since it's impractical to include all data of the individuals within the context of LLMs, memory modules have been incorporated into the agent framework, where relevant information is retrieved and provided to the LLMs. We use the BM25 algorithm (Robertson et al., 2009) to retrieve top-5 relevant records from the memory bank constructed in Sec. 3.2.

Role-playing Models Role-playing models are specially trained for role-playing specific characters, but can also be used to model untrained characters. The models either learn character knowledge from large-scale web corpus in pre-training or learn to role-play specific characters through fine-tuning (Yu et al., 2024; Lu et al., 2024).

In summary, for the nonparametric prompting methods, i.e., profile and profile with memory, we use vanilla LLMs described in the last section. For the parametric training-based schema, we include the existing open-source role-playing models **CharacterGLM-6B** (Zhou et al., 2023), **Ditto** (Lu et al., 2024) based on Llama-3-8B-Instruct, **Neeko** (Yu et al., 2024) based on Llama-3-8B, and close-source models **Baichuan-NPC-Turbo** [4] and **Xingchen-Plus** [5]. Since Ditto and Neeko have not released model weights, we used methods and data described in their papers for SFT.

4.4 Implementation Details

We set the temperature to 0.2 and limit the maximum token. For most questions, the limit is 32, except for the UR, where it is set to 128 to generate complete sentences. All input texts are formatted as conversations with consistent system messages, roles, and separators. The main experiment is conducted based on PoliSim, while the follow-up analysis experiments are based on the PoliSim-role subset. More details are in Appendix C.

5 Experiment Results

In this section, we evaluate the mainstream LLMs and analyze the experimental results.

[4] https://npc.baichuan-ai.com/index
[5] https://xingchen.aliyun.com/

5.1 Overall Results

Table 2 illustrates the results and we can find that:

- **How do different models perform?** Llama-3, GPT-3.5 and GPT-4 demonstrate the most impressive performance across dimensions. Some specialized role-playing models such as Ditto and Baichuan perform well in the cognition dimension but show less advantage in attitude and behavior. This may be because their training **emphasizes role knowledge and other dialogue system characteristics, without considering enhancements for these aspects**. CharacterGLM and Neeko tend to underperform compared to their general counterparts, i.e., ChatGLM-2 and Llama-3, since they are specialized for character dialogues but fall short in understanding and following instructions. While learning to simulate individuals, **maintaining general capabilities** is equally important.

- **How do different simulation schemes perform?** Models that integrate memory generally outperform those relying solely on profiles, particularly in attitude and behavior. This indicates that **both overall and detailed data about the individuals are important** for mimicking politicians. Besides, the degree of improvement from memory varies across models. Overall, the enhancement is more pronounced for smaller models, suggesting that memory mainly supplements the knowledge that models lack. Meanwhile, role-playing models show more pronounced advantages in certain abilities, such as the capacity to refuse to answer questions beyond one's cognitive boundary, which may be difficult to enhance through context alone.

- **What are the situations for different dimensions?** Most models are well-performed at RK of cognition dimension and IA and OA of attitude dimension since these questions can be answered with knowledge about the individual. However, most models are not aware to reject unknown questions without specific training. Also, the **simulation of behavior seems to be challenging** to the models. It shows an absence of capabilities to simulate individuals' thoughtful behaviors.

- **Are there certain politicians more difficult to simulate?** We analyze errors across politicians in Appendix D.2 and we find significant variation in the simulation results across politicians. The models perform poorly on moderate politicians and those whose decisions deviate from their party's mainstream positions. LLMs often **mispredict due to over-reliance on party affiliation as a dominant signal**.

Model	OA			EG		
	Acc.	Sim.	GPT Rank	Acc.	Sim.	GPT Rank
Llama-2-7B-chat-hf	83.7	86.8	2.7	45.8	86.2	3.0
Llama-3-8B-Instruct	**92.1**	87.9	2.2	**80.5**	87.5	2.8
Vicuna-7B-v1.5	84.0	85.3	3.9	71.0	81.9	3.2
Mistral-7B-Instruct-v0.3	90.0	**89.5**	**1.9**	71.2	**89.1**	**2.7**
ChatGLM2-6B	74.8	83.5	4.2	45.9	84.4	3.3

Table 4: Evaluation of OA and EG based on PoliSim multiple-choice questions, generated content similarity and LLM judgment by GPT.

5.2 Comparison with Generation-based Evaluation

Although multiple-choice questions provide an objective measure of model performance, they simplify the task to some extent. Thus, we conduct a generation-based evaluation for OA and EG to assess the consistency of results between the two evaluation methods. We calculate the cosine similarity of the generated response and the reference answer. We also follow Wang et al. to instruct GPT-3.5 to rank the responses and give explanations at the same time. As shown in Table 4, for open-source models, the rankings based on generation evaluations are generally consistent with those based on multiple-choice questions, except for the rankings between LLama-3 and Mistral. We prob into the samples and find it difficult to determine which model generates better even for human evaluators, making such evaluation potentially unstable.

5.3 Relationship between Dimensions

Cognitive behavioral theory suggests an interconnection between cognition, attitude, and behavior. We investigate whether this applies to LLM-based agents by using relevant QA pairs from the Cog., Att., and Beh. dimensions as historical dialogues to enhance responses in other dimensions, based on the *only profile* settings. As shown in Figure 3, cognitive information has minimal impact on attitudes and behaviors, likely because the data lacks highly relevant cognitive content. This suggests

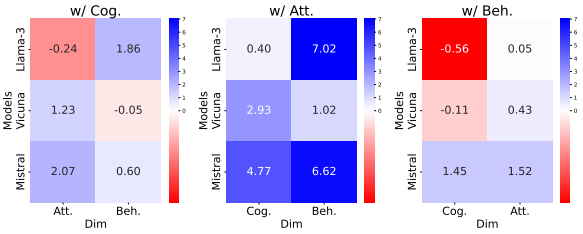

Figure 3: Improvement of performance after cognition, attitude and behavior enhancement through in-context learning. In the Cog. dimension, only RK is involved in the calculation since UR consists of irrelevant questions to the individuals, so it is unreasonable to retrieve related questions from the other two dimensions.

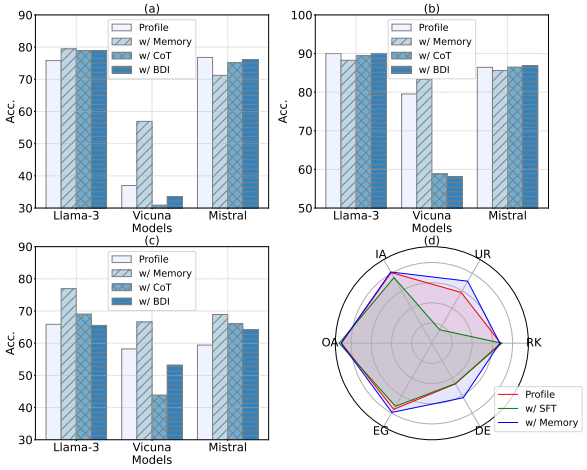

Figure 4: (a)-(c): Performance on Cog., Att. and Beh. under the profile settings, with memory, CoT and BDI. (d) Average performance of Llama3, Vicuna and Mistral on different dimensions under the profile settings with different knowledge augmentation methods.

that enhancing irrelevant cognition doesn't improve responses. In contrast, enhancing attitudes significantly and consistently influences the other dimensions. However, behavior's predictive power for attitudes is limited, possibly due to the insufficient detail in a single behavior record to fully capture ideological perspectives.

5.4 Leveraging Internal Reasoning Process

Besides relying on external components such as memory modules or additional training, we are curious about whether the simulation can be improved by leveraging LLMs' ability to model the internal thought processes of politicians. To investigate this, we explored two strategies: (1) **Chain of Thought (CoT)** (Wei et al., 2022): we prompt the agents to recall his or her personality, experiences, preferences and values first, and then answer. (2) **Belief-Desire-Intention (BDI)** (Rao and Georgeff, 1997; Adam and Gaudou, 2016): the BDI framework is a classical tool for modeling the decision-making process of politicians. We prompt the agents to write down his or her belief, desire and intention before answering. The details can be found in Appendix B. Figure 4 shows that these strategies can improve the performance of Llama-3 and Mistral, while hugely damaging the instruction-following ability of Vicuna, resulting in a substantial decrease in performance. It is the result of the fact that Vicuna can not generate effective reasoning processes as it often jumps out of the setting and claims to be an AI model. This further underscores that accurately simulating internal thought processes is the basis for modeling actual interactions.

5.5 Impact of the Form of Knowledge

Although existing work has explored providing role knowledge to models through external memory modules or parameter training, a comparative analysis of these approaches has yet to be conducted. We performed an experiment to determine whether providing the same information as memory within the prompt or integrating it through supervised fine-tuning (SFT) is more effective. Specifically, we convert the memory records retrieved into QA pairs using Llama-3. For example, an original statement about abortion could be transformed into a question asking for the individual's view on abortion and the response. We then use these dialogue data to fine-tune the models using LoRA (Hu et al., 2022). Figure 4(d) illustrates that although SFT may show improvements in certain dimensions such as RK and OA, it is prone to overfitting, which can lead to a significant decline in capabilities in UR. The memory-based method is more stable across dimensions, benefiting from the explicit and direct information provided through natural languages.

6 Conclusion

In this paper, we introduce PoliSim for evaluating LLM-based agents in politician simulation, grounded in cognitive behavioral theory. We evaluate agents across cognition, attitude, and behavior dimensions by re-formulating real-world data. Experiments reveal that strong base models with profiles and memory perform best, but behavior proves to be challenging for all models. To improve, modeling attitude can enhance other dimensions and effective modeling of internal processes is crucial for simulating politicians.

Limitations

Although we have constructed evaluation data across multiple dimensions as comprehensively as possible from real-world data, some limitations still remain:

- **Simplification of Politician Behaviors**: Currently, our focuses on evaluating mainstream simulation methods rather than proposing new models to capture the more complex internal processes of politicians. Internal strategic considerations and audience-focused reasoning may represent additional important dimensions worth investigation. However, obtaining comprehensive data on politicians' context, audience, and internal thought processes is challenging. Given these limitations, rather than modeling these unverifiable intermediate processes, we directly evaluate whether the behavioral outcomes align with those of the corresponding individuals.

- **Insufficient Interactive Evaluation**: Currently, except for the EG task, which includes group dynamics, all other evaluation tasks are single-turn, potentially missing simulation challenges unique to dynamic interactions. However, existing public data, whether from authoritative sources or political websites, primarily consist of **static** records without rich context. Apart from a very few politicians, such as presidents who have debate and other interactive data, it is challenging to obtain such data for most individuals. Thus, it is temporarily unavailable to objectively evaluate the consecutive behaviors of agents. Due to this condition, our benchmark still needs improvement. But we want to emphasize that given the current lack of even static, multi-dimensional evaluation frameworks, our work at least fills this gap.

- **Potential Bias in Data Construction**: We remind readers that using the proprietary model for dataset construction may introduce potential bias, which might make the tasks easier for the data generator model, i.e., GPT-3.5. This is a general systemic bias for benchmarks using model synthetic data. We will work on more data synthetic methods to minimize such risk.

Ethical Statement

We have minimized the ethical concerns as follows:

- **Data Collection and Privacy**: The data we use is completely available to the public and does not contain any information about privacy.

- **Benefit and Potential Misuse**: This paper aims to evaluate and improve LLMs' ability to simulate political actors, helping the public understand their representatives' positions and behaviors on key issues, and laying the foundation for applications such as policy simulation and event simulation. However, LLMs could be misused for other risky purposes, such as deepfakes, where features like speaking style are crucial. Therefore, we have avoided simulating these dimensions.

References

Carole Adam and Benoit Gaudou. 2016. Bdi agents in social simulations: a survey. *The Knowledge Engineering Review*, 31(3):207–238.

Alfred Adler. 2014. Individual psychology. In *An Introduction to Theories of Personality*, pages 83–105. Psychology Press.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Zachary R Baker and Zarif L Azher. 2024. Simulating the us senate: An llm-driven agent approach to modeling legislative behavior and bipartisanship. *arXiv preprint arXiv:2406.18702*.

Aaron T Beck. 1979. *Cognitive therapy and the emotional disorders*. Penguin.

Aditya Budhwar, Toshihiro Kuboi, Alex Dekhtyar, and Foaad Khosmood. 2018. predicting the vote using legislative speech. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, dg.o '18, New York, NY, USA. Association for Computing Machinery.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Xing Gao, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, Fei Huang, et al. 2024a. Roleinteract: Evaluating the social interaction of role-playing agents. *arXiv preprint arXiv:2403.13679*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024b. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520.

Joshua Clinton, Simon Jackman, and Douglas Rivers. 2004. The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370.

Shangbin Feng, Zhaoxuan Tan, Zilong Chen, Ningnan Wang, Peisheng Yu, Qinghua Zheng, Xiaojun Chang, and Minnan Luo. 2022. Par: Political actor representation learning with social context and expert knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12022–12036.

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. In *URL: https://data. stanford. edu/congress text*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Yupeng Gu, Yizhou Sun, Ning Jiang, Bingyu Wang, and Ting Chen. 2014. Topic-factorized ideal point estimation model for legislative voting network. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 183–192.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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 *International Conference on Learning Representations*.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023a. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2023b. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*.

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

Anastassia Kornilova, Daniel Argyle, and Vladimir Eidelman. 2018. Party matters: Enhancing legislative embeddings with author attributes for vote prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 510–515, Melbourne, Australia. Association for Computational Linguistics.

Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, et al. 2023. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv preprint arXiv:2310.02569*.

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. 2024. From skepticism to acceptance: Simulating the attitude dynamics toward fake news. *arXiv preprint arXiv:2403.09498*.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. Politics: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374.

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. *arXiv preprint arXiv:2401.12474*.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. 2024a. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*.

Xinyi Mou, Zejun Li, Hanjia Lyu, Jiebo Luo, and Zhongyu Wei. 2024b. Unifying local and global

knowledge: Empowering large language models as political experts with knowledge graphs. In *Proceedings of the ACM on Web Conference 2024*, pages 2603–2614.

Xinyi Mou, Zhongyu Wei, Lei Chen, Shangyi Ning, Yancheng He, Changjian Jiang, and Xuan-Jing Huang. 2021. Align voting behavior with public statements for legislator representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1236–1246.

Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024c. Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation. *arXiv preprint arXiv:2402.16333*.

Xinyi Mou, Zhongyu Wei, Qi Zhang, and Xuan-Jing Huang. 2023. Uppam: A unified pre-training architecture for political actor modeling based on language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11996–12012.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Anand S Rao and Michael P Georgeff. 1997. Modeling rational agents within a bdi-architecture. *Readings in agents*, pages 317–328.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023. Roleeval: A bilingual role evaluation benchmark for large language models. *arXiv preprint arXiv:2312.16132*.

Meta LLaMA Team. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

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

Keyon Vafa, Suresh Naidu, and David Blei. 2020. Text-based ideal points. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5345–5357.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

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*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Patrick Y Wu, Jonathan Nagler, Joshua A Tucker, and Solomon Messing. 2023. Large language models can be used to estimate the latent positions of politicians. *arXiv preprint arXiv:2303.12057*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey (2023). *URL https://arxiv. org/abs/2309.07864*.

Bushi Xiao, Ziyuan Yin, and Zixuan Shan. 2023. Simulating public administration crisis: A novel generative agent-based simulation system to lower technology barriers in social science research. *arXiv preprint arXiv:2311.06957*.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

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? *arXiv preprint arXiv:2404.12138*.

Yuqiao Yang, Xiaoqiang Lin, Geng Lin, Zengfeng Huang, Changjian Jiang, and Zhongyu Wei. 2021. Joint representation learning of legislator and legislation for roll call prediction. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1424–1430.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. 2024. Oasis: Open agents social interaction simulations on one million agents. *arXiv preprint arXiv:2411.11581*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, et al. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *arXiv preprint arXiv:2311.16832*.

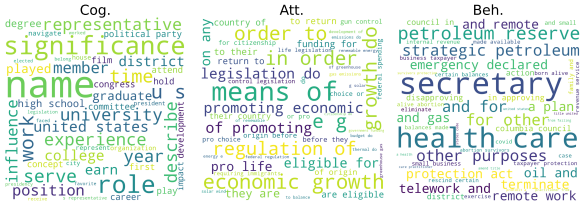

Figure 5: Word clouds of questions of different dimensions.

A Data

A.1 Memory Data

We collect public statements and voting records before 2023 from multiple sources, i.e., Twitter (Mou et al., 2023), VoteSmart, and Legiscan. In total, we have 3,045,530 records, 1,843,805 from tweets, 1,157,540 from voting records and 44,185 from statements.

A.2 Evaluation Data

Data Source For most dimensions, we use data after 2023 to construct evaluation questions in order to minimize data leakage as much as possible. Only for the EG dimension we use debate data from the 112th to 114th sessions, as this is the most recent data extracted from previous work (Gentzkow et al., 2018). We have supplemented the anonymization experiments to mitigate the impact of this data leakage.

Diversity Figure 5 shows the word clouds of the questions in different dimensions, showing that diverse topics have been involved.

B Prompt Details

B.1 Prompt for Evaluation Data Construction

Profile Generation We crawl the biographies of politicians from the website of congress and rephrase the structural information into natural languages using GPT-3.5.

Question Generation for Cognition Dimension We prompt GPT-3.5 to generate questions, answers and options from given wiki pages. The prompt for question generation is inspired by (Lu et al., 2024).

Prompt for Profile Generation

Given the following observation about {name}, please summarize the relevant details from the profile. His or her profile information is as follows:

Name: {name}
Profile Text: {text}
Party Affiliation: {party}
Represents State: {state}
Please avoid repeating the observations or sources of the content. Please directly summarize what kind of person this is:
Summary:

Prompt for Question Generation

You are skilled at designing questions for specific characters based on background information, as follows you will be provided with information for two characters:

[Character A] The name is {label1}, the description is {description1}. Here is an introduction to Character A: {wiki1}

[Character B] The name is {label2}, the description is {description2}. Here is an introduction to Character B: {wiki2}

Please design 3 questions that Character A can answer, but are not suitable for Character B to answer. The questions should strictly conform to Character A's era background and character setting, but go beyond the era, genre, occupation, age, knowledge, etc., settings of Character B, therefore Character B cannot answer them. Provide an explanation with each question, explaining why Character A can answer it but Character B cannot.

Please use as casual language as possible to ask questions, and try to use the second person for questioning, such as "Who are you?". Please respond in English. Please return the results in the following JSON structure:
[{{"question": str, "explanation": str}}]

Prompt for Response Generation

Please answer the questions according to your identity! When encountering questions that do not match your identity, please refuse to answer the question in the role of {label}, and explain the reason for refusal step by step based on your identity. Please do not step out of your role! Please avoid repeatedly restating your identity or name.

You are {label}, your description is {description}.
Here is your introduction:
{wiki}

Question: {question}

Prompt for Option Generation

You are a multiple-choice generator. Given a Character and his relevant knowledge, a question to the Character and his answer, you need to generate three additional incorrect option answers. The options need to go beyond the knowledge of the Character but keep the original tone and first person. Please ensure their plausibility and confusion with the original answer at the same time.
Character: {label}
Knowledge: {description} {wiki}
Question: {question}
Answer: {gt_response}
Please return the options in the format of:
Option1: option1
Option2: option2
Option3: option3

B.2 Question Generation for Attitude Dimension

To construct the questions for opinion alignment, we re-generate questions based on crawled statements:

Prompt for Question Generation of OA

Here is a piece of statement by {name}:
"{statement}"
Give a simple question to which {name} can respond with this statement.

B.3 Prompt for Question Answering in PoliSim

Here, we list the prompt used in the question answering in PoliSim. For *only profile* and *role-playing model* settings, we use the system prompt to assign the identity. For *profile+memory* setting, we retrieve and provide the memory through the user prompt. For multiple-choice questions, we follow (Li et al., 2023) to provide an unrelated question to guide the LLMs output in the required format. For the generation task, i.e., UR, we do not provide this sample.

Prompt for Question Answering without Memory

System prompt: You are {name}. {profile}
User prompt: Please answer the question and output your choice: Which of the following cities is located in the United States? Options: (A) New York; (B) Tokyo; (C) Beijing; (D) Paris.
Assistant Prompt: (A) New York
User prompt: Now please answer the question and output your choice: {question}

Prompt for Question Answering with Memory

System prompt: You are {name}. {profile}
User prompt: Please answer the question and output your choice: Which of the following cities is located in the United States? Options: (A) New York; (B) Tokyo; (C) Beijing; (D) Paris.
Assistant Prompt: (A) New York
User prompt: Your historical memory is: {memory}
Now please answer the question and output your choice: {question}

Model	Cog.		Att.		Beh.	
	Acc.	w/ anon. Acc.	Acc.	w/ anon. Acc.	Acc.	w/ anon. Acc.
Only Profile						
Llama-2	52.0	42.7	77.3	77.3	43.5	42.0
Llama-3	75.8	75.4	90.0	89.9	65.9	68.8
Vicuna	37.0	40.5	79.5	79.3	58.2	57.5
Mistral	76.8	67.7	86.4	87.8	59.4	58.3
ChatGLM-2	26.4	27.6	72.4	72.3	52.2	50.5
Profile+Memory						
Llama-2	54.4	43.0	74.7	80.4	54.6	51.9
Llama-3	79.5	74.5	88.3	90.0	77.0	75.6
Vicuna	56.9	47.3	83.3	83.7	66.6	64.7
Mistral	71.2	62.7	85.6	87.9	68.9	69.5
ChatGLM-2	40.9	38.1	72.5	73.2	51.7	51.0

Table 5: Performance of open-source models with and without anonymization strategies.

B.4 Prompt with CoT / BDI

Prompt with CoT

Read the question and write your personality, experience, preferences and values.
Question: {question}

Prompt with BDI

Read the question and write your BELIEF, DESIRE and INTENTION.
Question: {question}

C Implementation Details

Our benchmark and the evaluation framework are PyTorch-based. All experiments are conducted on 8 NVIDIA GeForce RTX 4090 24GB GPUs. During the evaluation, half precision is used to accelerate the process. Since the cost of LLM inference can be very high and our data size is larger than previous work (Wang et al., 2023b; Shen et al., 2023; Chen et al., 2024a), we only run once for each model. We follow Li et al. to prepend an in-context sample to guide the models to generate responses in the desired format and shuffle the options to reduce bias brought by option marks.

D Supplementary Experiment

D.1 Data Leakage and Anonymization

Compared to previous character data based on fictional scripts, even though we adopt a better strategy by constructing recent data for testing purposes, there is still a risk of data leakage. The data used might have already appeared in the model's pre-training data. To address this, we adopt an entity replacement strategy (Xu et al., 2024), using *[Character A]* to anonymize the individuals. Table 5 indicates that anonymization has the greatest impact on the simulation of cognition. It is reasonable, as assessing a person's knowledge and cognitive boundaries must be based on his or her real identity, while attitudes and behaviors can be partly inferred through the reading comprehension of the profile.

Model	Cog.		Att.		Beh.	
	RK	UR	IA	OA	EG	DE
Llama-2-7B-chat-hf	26.97	25.13	14.76	11.37	30.77	43.95
Llama-3-8B-Instruct	27.04	19.17	10.54	6.96	23.33	38.38
Vicuna-7B-v1.5	28.15	24.01	18.20	11.44	27.81	38.93
Mistral-7B-Instruct-v0.3	25.61	21.65	15.17	10.59	27.86	38.44
ChatGLM2-6B	26.10	29.02	12.62	14.60	28.52	16.15
GPT-3.5-Turbo	25.22	23.97	13.59	9.84	23.02	27.80
GPT-4-Turbo	24.41	18.09	13.49	7.18	16.82	18.42

Table 6: std of performance across politicians in the simulation of "only profile" settings.

D.2 Performance across Politicians

In Table 6, we calculate the variance in performance at the role-level in Table 3 and analyze specific cases. By doing this, we found significant variation in the simulation results across politicians. For some moderate politicians (as inferred from their ideological positions on GovTrack), such as McHenry, Patrick T., Scalise, Steve, and Bishop, Sanford D, the models achieved the lowest accuracy on more than half of the dimensions. Additionally, there are some politicians who make choices on certain issues that are not aligned with the mainstream of their party. LLMs tend to rely heavily on the strong signal of party affiliation, which leads to prediction errors in these cases.