How Far Are LLMs from Believable AI? A Benchmark for Evaluating the Believability of Human Behavior Simulation

Anonymous ACL submission

Abstract

In recent years, AI has demonstrated remarkable capabilities in simulating human behaviors, particularly those implemented with large language models (LLMs). However, due to the lack of systematic evaluation of LLMs' simulated behaviors, the *believability* of LLMs among humans remains ambiguous, i.e., it is unclear what LLMs' level of believability is. In this work, we design *SimulateBench* to evaluate the believability of LLMs when simulating human behaviors. In specific, we evaluate the believability of LLMs based on two critical dimensions: 1) *consistency*: the extent to which LLMs can behave consistently with the given information of a human to simulate; and 2) *robustness*: the ability of LLMs' simulated behaviors to remain robust when faced with perturbations. SimulateBench includes 65 character profiles and a total of 8,400 questions to examine LLMs' simulated behaviors. Based on SimulateBench, we evaluate the performances of 10 widely used LLMs when simulating characters. The experimental results reveal that current LLMs struggle to align their behaviors with assigned characters and are vulnerable to perturbations in certain factors. [1]

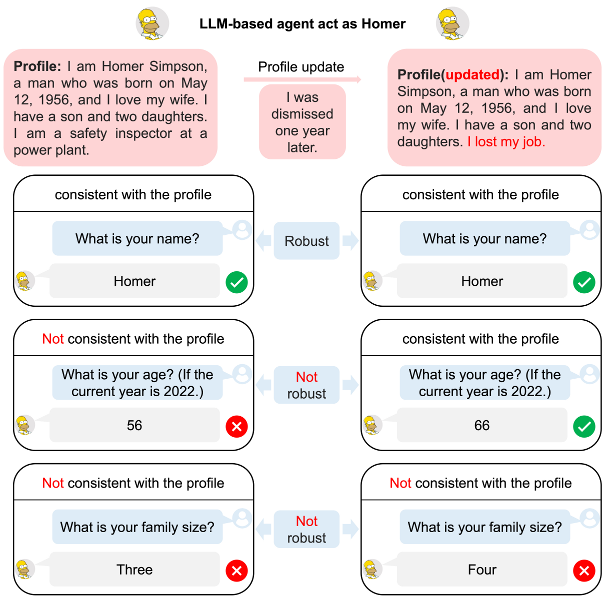

Figure 1: An illustrative example of the "Consistency", and "Robustness". Consistency measures whether the LLMs' generated human behavior accurately depicts the profile information; Robustness measures whether the generated human behavior will be influenced by the perturbation in the profile.

1 Introduction

AI has shown promise to simulate human behavior and social interaction (Wooldridge and Jennings, 1995; Macal and North, 2005), which can empower applications ranging across prototyping social theories (Aher et al., 2023; Horton, 2023; Kovač et al., 2023), generating synthetic research data (Hämäläinen et al., 2023; Wang et al., 2023a) and building non-player characters (Laird and VanLent, 2001). These applications necessitate the simulated human behavior to possess a convincing level of *believability*, which allows the users to suspend their disbelief (Ortony et al., 2003). Such believability is crucial as it facilitates users in establishing trust in the AI and streamlines the fulfillment of the AI's goals in these applications.

Despite the importance of believability, the current believability level of LLMs remains unclear. Previous studies have primarily assessed believability using human ratings, GPT-based evaluations, or case studies (Park et al., 2022, 2023; Argyle et al., 2023; Hämäläinen et al., 2023). While these approaches provide valuable insights, they are not without limitations. Such evaluations often suffer from inter-task inconsistency and are susceptible to biases introduced by either human evaluators or the models themselves. To address these challenges, this paper introduces a systematic method

[1] Code and SimulateBench are available at an anonymous GitHub repository.

for evaluating the believability of LLM simulations. Specifically, we focus on improving the evaluation of consistency and robustness, as illustrated in Figure 1. Consistency means that the behaviors of LLMs must align with the character's characteristics. Breaking this consistency will cause disbelief (Loyall, 1997). Robustness requires the LLMs to maintain the same behaviors when nuanced updates and modifications, denoted as perturbations, are performed on the input.

To this end, we propose evaluating the believability of LLMs by (1) consistency: To what extent does the generated human behavior accurately depict the profile? (2) robustness: To what extent do the LLMs' behaviors maintain robustness when faced with perturbations in the profile? To measure consistency and robustness, we introduce SimulateBench, a benchmark for character data collection and evaluation of consistency and robustness. SimulateBench consists of four parts: the profile descriptive framework, the character profile dataset, the consistency dataset, and the robustness dataset. The profile descriptive framework is proposed to guide annotators in comprehensively documenting a character's profile: sufficient profile information will ensure more accurate and effective simulations, which also align with real-world application scenarios. Based on the framework, we collect a character profile dataset, including the profiles of 65 characters. To measure the consistency, we assess whether the LLMs can correctly answer multi-choice questions about the character in the consistency dataset. To correctly answer these questions, the LLMs must participate in logical reasoning based on the profile. To measure the robustness, we perturb the profiles in the consistency dataset to construct the robustness dataset and compare how the LLMs' consistency ability changes.

Through the SimulateBench, we evaluate the level of believability of ten widely used LLMs. Our findings show that 1) LLMs perform poorly for consistency: they can not accurately depict the information in the comprehensive profile input, even if they are equipped with long context size; 2) LLMs exhibit a lack of robustness when faced with even nuanced profile perturbation; 3) LLMs exhibit bias towards some perturbations. In further studies, we examine four influential factors that will greatly influence the LLMs' believability.

In summary, we propose two novel dimensions of consistency and robustness to measure LLMs' believability. To facilitate the assessment, we introduce the SimulateBench. We hope our work will inspire further research into the believability of human behavior simulation.

2 Related Work

2.1 Human behavior Simulation

Recently, LLMs have demonstrated intelligence comparable to humans in certain tasks (bench authors, 2023; Brown et al., 2020; Touvron et al., 2023). Many studies endeavor to harness the LLMs to simulate human behavior and social interactions in social science, economics, psychology, and human-computer interaction for prototyping theories and generating synthetic research data (Park et al., 2022, 2023; Argyle et al., 2023; Horton, 2023; Hämäläinen et al., 2023). Other studies prompt LMs(LLMs) with profiles to simulate human conversations in role-playing and personalized dialogue (Zhang et al., 2018; Zheng et al., 2019, 2020; Wang et al., 2023b; Chen et al., 2023). However, their provided profile to LLMs is concise, which is far from real scenarios. The limited amount of personal information provided is insufficient for the model to acquire sufficient knowledge to simulate a character accurately. Therefore, we propose collecting a comprehensive character profile to meet the demand of real-world application scenarios.

2.2 Evaluation of LLMs in Human Behavior Simulation

Simulation of human behavior requires the LLMs to faithfully embody assigned roles and identities and proactively interact with others (Wooldridge and Jennings, 1995; Franklin and Graesser, 1996; Ortony et al., 2003). See et al. (2019); Fang et al. (2023); Choi et al. (2023) propose evaluation frameworks toward LLMs' capabilities of natural language understanding and generation. Rao et al. (2023); Jiang et al. (2023); Huang et al. (2023) evaluate LLMs' abilities to understand and maintain personality traits. Aher et al. (2023) introduce the Turing Experiment to assess whether or not LLMs can simulate the behavior of a representative sample of participants in human subject research. Park et al. (2023) propose a sandbox and an online social network to evaluate agents' interactions. Ahn et al. (2024) proposes evaluating LLMs when role-playing at a specific time. However, little research assesses the LLMs' level of believability in consistency and robustness in real scenarios where a comprehensive profile is provided. Hence, we aim to bridge this gap by constructing SimulateBench.

3 SimulateBench

We introduce SimulateBench for character profile collection and believability evaluation. Specifically, our benchmark includes the profiles of 65 characters and 8400 questions to assess the LLMs' consistency and robustness when simulating human behavior. The statistics are shown in Table 1.

Statistical categories	Number
Characters	65
Avg tokens per profile	3277
Avg tokens per question	58
Avg questions per character	#
Immutable Characteristic	41
Social Role	52
Relationship	57
Total benchmark questions	8400

Table 1: The statistics of SimulateBench. The tokens are counted with the tokenizer of GPT-4.

3.1 Profile Descriptive Framework and Character Dataset

Comprehensive profile information is necessary for LLMs to simulate human behavior accurately. Accordingly, we propose the profile descriptive framework and collect a character dataset based on this framework. For more details, please refer to the Appendix A.

Profile Descriptive Framework We propose a descriptive framework that comprehensively documents a character's profile from three attributes: **Immutable Characteristic**, **Social Role**, **Relationship**. Immutable characteristic (Stein, 2001) refers to characteristics that cannot be easily changed, such as name, gender, and age. Social role (Wasserman, 1994; Eagly and Wood, 2012) is conceptualized as a set of connected behaviors, obligations, beliefs, and norms as conceptualized by people in a social situation. Relationship (Sztompka, 2002) is the basic element of study in the field of social sciences and refers to any interpersonal connection between two or more individuals. Furthermore, these three kinds of profile information are thoroughly elaborated by fine-grained aspects based on established theories. For example, we will comprehensively document the following attributes of the relationship: **familiarity**, **judgment**, **affection**, **behavioral patterns**, **relationship status**, and **communication history**. The annotators will collect the profiles according to the attributes defined by the framework.

Character Dataset We selected characters from TV dramas of popular genres[2]: The Simpsons (Animated), Friends (Comedy), Breaking Bad (Crime), and The Rings of Power (Science fiction). We do not collect real human profiles for ethical reasons, such as preventing information leaks. According to the profile descriptive framework, annotators extract the profile information from the fandom[3]: a wiki hosting service that hosts wikis mainly on entertainment characters. We recruited four PhD students to collect the profile information from fandom based on the profile framework. First, we ask one annotator to collect the characters' profiles. Then, we asked the remaining three annotators to review the collected data. If there are disagreements among the three reviewers, all four annotators will discuss and modify or remove the collected information. Through this process, 5.67% of the profile tokens are modified or removed. We will leave it blank if there is no content about one attribute. Finally, the resulting profiles were stored in JSON format: *{attribute of the profile: corresponding content}*. As shown in Table 1, every profile contains 3,277 tokens on average, which is comprehensive in comparison to prior studies. As an illustration, the profile mentioned in the well-known study by Park et al. (2023) only contains 203 tokens.

3.2 Measuring Consistency

Consistency Dataset The consistency dataset is composed of multi-choice questions. Each character has an average of 150 questions. To answer these questions accurately, the LLMs need to analyze and employ logical reasoning to the profile information.

Question We will design a template question for every attribute in the profile descriptive framework. Then, we apply these template questions to each character to generate the corresponding questions. Figure 2 shows an example of this process.

Options and Ground Truth For every question related to one profile attribute, we extract the cor-

[2] https://www.imdb.com/list/ls023983860/

[3] https://www.fandom.com/

responding content of this attribute as the ground truth of this question from the JSON-formatted profile. We add an option of "There's not enough information to answer this question.". This option is intended for the blank attribute in the profile, and we set this option as the gold answer in such a case. The reason for this setting is that if the LLM is given unrestricted freedom to respond to the content that is not mentioned in the profile, there is a high probability of compromising the character's information and undermining the LLM's believability. We categorize the questions into two classes according to their gold answer: **Known** and **Unknown**. Unknown's gold answer is "There's not enough information to answer this question".

Validation We ask the four annotators to validate the quality of the question, options, and ground truth. If the ground truth is misaligned with the question and the profile, the annotators will discuss and then remove or modify this question and corresponding options and ground truth. Finally, 7.18% of questions are removed or modified.

Measuring Metric: CA To measure the consistency, we will employ the LLMs to answer the questions in the consistency dataset, and we will calculate the accuracy of these answers as the consistency ability, referred to as *CA*.

3.3 Measuring Robustness

Robustness Dataset The robustness dataset is constructed by perturbing the characters' profiles (denoted by the characters' variant) and modifying the questions in the consistency dataset accordingly. We perturb the profile of characters by replacing the content of demographic attributes: **Education**, **Surname**, **Race**, and **Age**. To prevent irrationality caused by the perturbation, a thorough examination of the consequences resulting from any modifications made to the initial profile is conducted. According to this perturbation, we modify the corresponding questions in the consistency dataset. Then, we include the modified questions in our robustness dataset. For instance, if we modify the age of a character from 20 to 30, our initial step will involve duplicating the questions pertaining to the character in the consistency dataset. Subsequently, we shall alter these questions and their gold answers to align with the age adjustment. After the alteration of these questions, we get the questions for the character at the age of 30.

Question template
Attribute: Age(Birth year)
Question: What is your age group? (We are in the calendar year 2024)

Options: A. Under 18; B.18-24; C.25-34; D.35-44; E.45-54; F.55-64; G.65 or above; H. There's not enough information to answer this question;

Process to get the ground truth of the character Homer
The annotators first extract the attribute content of the birth year of Homer from its profile, which is 1956. Then, the annotator calculates the age: 2024-1956=68. So, the ground truth is G. 65 or above.

Ground truth: G.65 or above.

Figure 2: An illustrative example of the template question and the process to get the ground truth.

Measuring Metrics: RA and RCoV The robustness aims to determine the variation in the consistency performance of the LLMs when slight perturbations are made to profiles. To achieve this goal, we employ the standard deviation of CA and coefficient of variation[4] of CA as the robustness performance of LLMs, referred to as *RA* and *RCoV* respectively. For example, when employing GPT-4 to simulate a character, only modifying the age attribute in the profile to values of 10, 15, 20, 25, and 30 yields five variants. After all five variants answer the questions in the corresponding robustness dataset, five CA scores will exist: $s_1, \ldots, s_5$. The five scores' standard deviation and mean are σ and μ, respectively. The RA of GPT-4 will be σ. The RCoV of GPT-4 will be σ/μ.

Dividing RA by μ allows for the comparison of different models. RCoV can be understood as the quantification of the impact that robustness (RA) can have on the actual performance (μ). As an illustration, LLM A demonstrates an RA of 0.04, a μ of 0.3, and hence RCoV to be 0.13. LLM B exhibits an RA of 0.08, a μ of 0.9, and hence RCoV to be 0.089. While LLM B has a higher RA score (0.08 compared to 0.04), the actual impact of its RA on performance is smaller (0.089 compared to 0.13).

4 Baseline Methods for Human Behavior Simulation

Three components are crucial to prompting the LLM to simulate human behavior: the instruction to explain how to simulate human behavior (I), the profile of specific characters (II), and the description of the task (III). Below, we introduce how we

[4] https://en.wikipedia.org/wiki/Coefficient_of_variation

Model	CA	Immutable Characteristic		Social Role		Relationship	
		Known	Unknown	Known	Unknown	Known	Unknown
GPT-4	0.77	1.00	0.47	1.00	0.59	0.97	0.06
Qwen2.5-7B	0.73	0.91	0.42	0.97	0.59	0.91	0.13
GPT-3.5	0.70	0.82	0.58	0.56	0.88	0.91	0.31
XVERSE-13B	0.62	0.68	0.53	0.68	0.76	0.59	0.44
Vicuna-13B	0.59	0.64	0.32	0.76	0.18	0.76	0.56
ChatGLM2-6B-32K	0.55	0.68	0.21	0.71	0.24	0.79	0.25
ChatGLM2-6B	0.49	0.50	0.16	0.65	0.12	0.88	0.06
Qwen2.5-3B	0.48	0.41	0.84	0.38	0.94	0.03	0.81
Vicuna-7B	0.46	0.36	0.05	0.85	0.06	0.74	0.06
Llama-3.1-8B	0.10	0.36	0.00	0.09	0.00	0.09	0.00
Average	0.55	0.64	0.36	0.66	0.44	0.67	0.27

Table 2: CA scores across ten models to simulate a character. The last six columns correspond to the accuracy of the model for different types of questions. A larger CA indicates better consistency performance.

implement these three components in our baselines.

I: Simulate Human Behavior For models like GPT-4 that have gone through RLHF (Wirth et al., 2017; Stiennon et al., 2020), the RLHF will equip LLMs with specific language preferences and habits, such as introducing itself "as a language model", which will harm the believability. To overcome these issues, we set an *instruction prompt template* to instruct the LLM on how to simulate human behavior.

II: Profile of Specific Characters we will fill in the collected profile of the character in the *instruction prompt template* to incorporate the knowledge about the character into LLMs.

III: Prompting for Consistency Dataset Given that our assessment of consistency is performed in a question-answering format, the prompt for the task is: *Answer the below question; you should only choose an option as the answer. Choose "I do not know" if there is insufficient information to answer the question. {example}. {question}* . The placeholder of *{example}* will be filled if few-shot (Brown et al., 2020) is applied in the experiments. Additionally, chain-of-thought (CoT) (Wei et al., 2022) and Self-Ask (Press et al., 2022) will be utilized in zero-shot and few-shot settings. In summary, five combinations of prompting strategies and learning settings are considered: **Zero**, **Zero+CoT**, **Few**, **Few+CoT**, **Few+Self-Ask**.

III: Prompting for Robustness Dataset The prompting used for the robustness dataset is similar to the one for the consistency dataset. The difference lies in that we will prompt the perturbed profile of the character to the instruction prompt template. In this way, the LLM can simulate the character's variants, and we will compute the RA and RCoV when the LLM simulates these variants to evaluate the robustness of the LLM.

5 Experiment

5.1 Experimental Setup

We comprehensively assess 10 LLMs, including commercial models and open-source models. Among these models, GPT-3.5 and GPT-4 are commercial models, and other models are open-sourced models. We access the open-source LLMs from their official repositories in Hugging Face[5]. We use a fixed version of the above models and set the temperature to 0 to help reproducibility.

5.2 Consistency Evaluation Results

Table 2 shows various models' CA scores across all question types when simulating a character. We have the following findings:

GPT series perform better than open-source models; longer context size does not necessarily mean better consistency performance For GPT-4 and GPT-3.5, the CA scores across six question types are 0.77 and 0.7, respectively. In comparison, the open-source models perform worse, with the lowest average CA of Llama-3.1-8B being 0.1. This observation highlights a significant disparity between open-source and GPT series models. In some studies (Qian et al., 2023; Park et al., 2023), it is observed that the decision-making processes highly rely on the GPT-3.5, which is expensive compared to open-source models. When researchers want to use an open-source model as a

[5] https://huggingface.co/

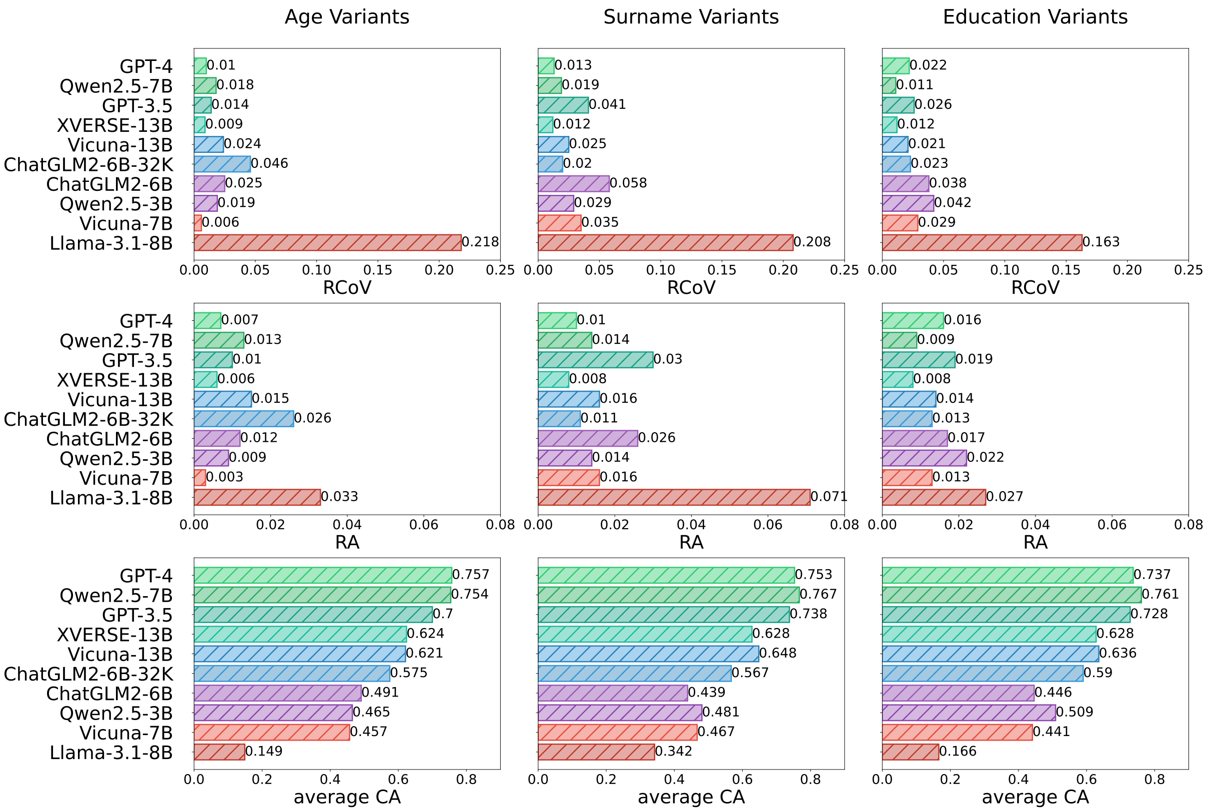

Figure 3: The RCoV, RA, and CA scores of models to simulate the variants of a character. A smaller RCoV indicates stronger robustness, while a larger CA indicates stronger consistency.

substitute to reduce expenses and enhance usability (Kaiya et al., 2023), it is crucial to consider this disparity.

Furthermore, although equipped with a longer context size of 128k, the performance of the Llama-3.1-8B is worse than the GPT-4(8K) and ChatGLM2-6B(8K). This implies that increasing the context window size does not necessarily result in improved consistency performance.

Models demonstrate severe simulation hallucination As seen by the data presented in the table 2, it is apparent that the accuracy for Unknown questions is considerably lower than that of the known questions. Even the best GPT-4 performs worse, with a CA score of 0.06 for the unknown relationship questions. This observation indicates that when the available information in the profile is insufficient to address the query, these models tend to provide nonsensical responses rather than adhering to the prescribed instruction, which requires the LLMs to answer with "I do not know" in such a case. This greatly undermines the credibility of the models. For example, when GPT-3.5 acts as Homer and is questioned about his religious convictions, its response indicates Christian. Nevertheless, the profile provides no evidence of Homer's adherence to Christianity. The model may deduce Homer's religious views just by Homer's Caucasian ethnicity. Inspired by the definition of hallucination (Zhang et al., 2023), we refer to the phenomenon as simulation hallucination.

5.3 Robustness Evaluation Results

The results are shown in Figure 3. The RCoV, RA, and CA scores are reported when models are instructed to simulate a character and perturbations are conducted on the character's profile. The finding is:

Better consistency performance does not necessarily mean better robustness performance As shown in Figure 3, models that exhibit strong consistency performance may yet demonstrate inadequate robustness performance. For instance, Vicuna-13B(0.621) outperforms Vicuna-7B(0.457) in terms of consistency in the Age Variants group, but Vicuna-13B exhibits worse robustness(RCoV of 0.024 larger than 0.006 of Vicuna-7B; RA of 0.015 larger than 0.003 of Vicuna-7B). Only the GPT series has a relatively high level of both consistency and robustness. This indicates that LLMs also face challenges in terms of robustness.

Open-source models show poor robustness; models exhibit the same level of robustness towards different perturbations Some open-source models show poor robustness when faced with profile perturbation. For example, the Llama-3.1-8B model exhibits severe performance, reaching a 0.218 RCoV score and a 0.033 RA score in the Age Variants group; 0.218 RCoV score indicates that perturbations can impact the model's consistency performance up to 21.8%.

Moreover, The RCoV and RA scores for all three variants also revealed that the model will demonstrate similar robustness performance even when faced with different perturbations, as shown in Table 3. That means that the models show relatively the same level of robustness towards different perturbations. That means that the models' robustness level may be an inherent property that is not influenced by the perturbation types.

6 Influential Factors for Believability

This section delves deeper into the four factors that exert substantial influences on believability. We anticipate that our studies could expedite subsequent research on human behavior simulation.

Simulation hallucination As shown in Table 2, models demonstrate severe simulation hallucination with CA of Unknown questions is considerably lower than that of Known questions. One plausible possible explanation is that the model might have known the answer to a question due to the knowledge learned in the training process, even if the answer can not be deduced from the profile. Consequently, the model refuses to answer the question with "I do not know." as required in the prompt [6]. This phenomenon reflects that models occasionally prefer to refuse or ignore the user's instructions, which will greatly harm the user's believability towards the model, especially when commercial simulation products are gaining increasing popularity, such as character.ai and npc.baichuan-ai.

Bias of models towards specific demographic attributes We have found that believability can be significantly influenced by the profile perturbation in Section 5.3. Hence, it is crucial to determine which profile information would yield high

[6] In Appendix D, we further examine the effect of simulation hallucination by replacing the name of the character to compare the variants' CA scores of Unknown questions.

Variant Pair	Age & Education	Age & Surname	Education & Surname
RCoV	**0.96**	**0.96**	**0.98**
RA	0.47	**0.66**	**0.76**

Table 3: The correlation coefficient of models' RCoV and RA scores of variant pairs. Bold indicates that the results are significant with $p < 0.01$.

Age	1956	1985	2000
Average CA	0.63	0.60	**0.65**
Name	Keams	Bedonie	Nguyen
Average CA	0.64	**0.69**	0.61
Education	High School	Middle School	Bachelor
Average CA	0.64	0.62	**0.69**
Race	African	Caucasian	Middle Eastern
Average CA	0.60	**0.63**	0.56

Table 4: The average CA scores of known questions of models when simulating the variants of a character.

believability for various LLMs. To investigate this question, we compare the LLMs' consistency by perturbing different demographic attributes in the profile. Specifically, we employ LLMs to simulate Homer by prompting the profile of Homer's variants in the character variants dataset, whose profile is modified with only one demographic attribute, such as birth year, while keeping all others unaltered.

Table 4 shows the results. All LLMs exhibit various degrees of preference toward profiles with specific demographic attributes. Models exhibit a significantly higher consistency score for the race of Caucasian (0.63) over the Middle Eastern (0.56), the education of bachelor (0.69) over the middle school, the name of Bedonie (0.69) over Nguyen, and birth year of 2000 (0.65) over 1985 (0.60). This observation indicates that models consistently prefer specific demographic attributes. This phenomenon may be attributed to the fact that models are trained on overlapping corpora, resulting in the corpus bias being simultaneously manifested in all these models.

Position in the profile For long textual inputs, models can pay different attention to the information in different positions. Hence, the believability can be impacted by the placement of information inside the profile. To investigate this issue, we conduct experiments by adjusting the order of infor-

Model	Known		Unknown	
	Normal	*Reverse*	*Normal*	*Reverse*
GPT-4	1.00	0.95	0.47	0.47
Qwen2.5-7B	0.91	0.91	0.42	0.53
GPT-3.5	0.82	0.77	0.58	0.63
ChatGLM2-6B-32K	0.68	0.73	0.21	0.32
XVERSE-13B	0.68	0.73	0.53	0.53
Vicuna-13B	0.64	0.68	0.32	0.37
ChatGLM2-6B	0.50	0.59	0.16	0.32
Qwen2.5-3B	0.41	0.42	0.84	0.84
Vicuna-7B	0.36	0.64	0.05	0.11
Llama-3.1-8B	0.36	0.56	0.00	0.00
Average	0.64	**0.70**	0.36	**0.41**

Table 5: The accuracy of Immutable Characteristic questions for models to simulate a character with the profile's information order reversed (denoted as *Reverse*) and unchanged (denoted as *Normal*).

mation in the profile. The original profile presents information in the order of Immutable Characteristic, Social Role, and Relationship, indicated as *Normal*. The adjusted order, denoted as *Reverse*, is Social Role, Relationship, and Immutable Characteristic. Then, we evaluate LLMs through the consistency dataset.

Table 5 shows the results. The revised sequence order has significantly improved the CA scores of open-source models on the Immutable Characteristic questions: the average CA of reverse known questions is 0.7 compared with the normal of 0.64, and the average CA of reverse unknown questions is 0.41 compared with the normal of 0.36. Nevertheless, this effect is not apparent for the commercial models. A possible explanation is that open-source models may struggle to adequately process lengthy textual content, even when their context size is large enough. Consequently, the model will allocate different attention to the information in the prompt's different positions. Nevertheless, the commercial models retain strong processing capabilities when it comes to handling lengthy texts. Therefore, altering the sequence order is less likely to significantly influence the commercial model's performance.

Reasoning prompting Although reasoning prompting techniques, such as chain-of-thought, are considered effective in some tasks, we find they can not always increase the believability of human behavior simulation. To provide evidence, we conduct the simulation using prompt combinations of Few, Few+CoT, Few+Self-Ask, Zero, and Zero+CoT.

Model	Few	Few+ CoT	Few+ Self-Ask	Zero	Zero+ CoT
GPT-4	0.77	0.77	**0.82**	0.75	0.77
Qwen2.5-7B	**0.73**	**0.73**	0.71	0.59	0.51
GPT-3.5	0.70	0.77	0.77	0.77	0.77
XVERSE-13B	**0.62**	0.42	0.43	0.60	0.58
Vicuna-13B	0.59	0.61	0.63	**0.65**	**0.65**
ChatGLM2-6B-32K	0.55	**0.63**	0.59	0.58	0.58
ChatGLM2-6B	0.49	**0.54**	0.49	0.44	0.41
Qwen2.5-3B	**0.48**	0.46	0.42	0.47	0.46
Vicuna-7B	0.46	0.54	0.56	**0.58**	**0.58**
Llama-3.1-8B	0.10	0.04	0.06	**0.13**	0.11

Table 6: : The CA scores of models when simulating Homer with five different prompting strategies.

Table 6 shows the results. Among all the prompt combinations considered, it is seen that no prompt combination exhibits a consistent improvement in the performance of all the models when compared to other prompts. One plausible explanation posits that the efficacy of these prompt techniques, such as CoT and Self-Ask, primarily lies in their ability to enhance performance on tasks involving reasoning abilities, such as solving, decision-making, and planning (Huang and Chang, 2022; Wang et al., 2022). Nevertheless, simulating human behaviors necessitates the model to hold other abilities, such as comprehensive comprehension of the character's profile and the dynamics of character relationships.

We also find that some open-source models, such as the Vicuna series, perform even better when no demonstration examples are included in the prompt (Zero) compared with the Few setting. We carefully analyzed their responses and found that these models consistently generate the exemplars in the Few setting as a response. One potential reason is that the lengthy profile and the challenging task complexity hinder the model from comprehending the exemplar in the Few setting.

7 Conclusion

We proposed two novel dimensions to measure LLMs' level of believability: consistency and robustness. We introduced SimulateBench, a benchmark for the profile collection and measuring LLMs' consistency and robustness. Through the SimulateBench, we evaluated the level of believability of popular LLMs. Our experimental results and findings provided insights to facilitate future research on developing human-like AI.

Limitations

In this paper, we proposed two dimensions to measure LLMs' level of believability when simulating human behavior. Simulating human behavior is an intricate undertaking that necessitates extensive and detailed information on the character's profile. Despite the fact that our work has a considerably thorough profile compared to earlier works, it may still be inadequate. Furthermore, despite our thorough evaluation of many well-known models, certain commercial models, such as Claude from Anthropic, have not been included in our evaluation. This omission is due to the requirement of qualification audits for using these models, which we do not have access to. Consequently, the evaluation of these models is not included in our research.

Ethics Statement

Annotators and contents We strictly adhere to the ACL Code of Ethics. We placed high importance on ensuring the comfort and well-being of our annotators. We advised them to stop the annotation process if they came across any information that caused them discomfort. We recruited annotators at a rate of $2 \sim 3$ times their local hourly minimum wage. We instruct the annotators to collect data without bias and keep the content free from unsafe, toxic, biased, offensive, and harmful content. We utilize the models in accordance with their designated purpose. In summary, we make every effort to adhere to the ethical norms set forth by ACL.

Anthropomorphism Simulation is a technique that allows large language models (LLMs) to simulate human-like behavior to fulfill user requirements. Although assessing the simulation capabilities of LLMs via our benchmark may prompt anthropomorphic interpretations-assigning human-like attributes to LLMs-it is crucial to underscore that our objective is not to humanize LLMs. Our purpose is to augment the capacity of LLMs to simulate human behavior, hence enhancing human-machine interaction. This initiative aims to bridge the interaction divide between humans and machines, while acknowledging the essential characteristics that distinguish them.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

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 *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3291–3325, Bangkok, Thailand. Association for Computational Linguistics.

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.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. *arXiv preprint arXiv:2305.14938*.

Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. 1997. Human conversational behavior. *Human nature*, 8:231–246.

Alice H Eagly and Wendy Wood. 2012. Social role theory. *Handbook of theories of social psychology*, 2:458–476.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Stan Franklin and Art Graesser. 1996. Is it an agent, or just a program?: A taxonomy for autonomous agents. In *International workshop on agent theories, architectures, and languages*, pages 21–35. Springer.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023. Peacok: Persona commonsense knowledge for consistent and engaging narratives. *arXiv preprint arXiv:2305.02364*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2023. Personallm: Investigating the ability of gpt-3.5 to express personality traits and gender differences. *arXiv preprint arXiv:2305.02547*.

Zhao Kaiya, Michelangelo Naim, Jovana Kondic, Manuel Cortes, Jiaxin Ge, Shuying Luo, Guangyu Robert Yang, and Andrew Ahn. 2023. Lyfe agents: Generative agents for low-cost real-time social interactions. *arXiv preprint arXiv:2310.02172*.

Grgur Kovač, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. The socialai school: Insights from developmental psychology towards artificial socio-cultural agents. *arXiv preprint arXiv:2307.07871*.

John Laird and Michael VanLent. 2001. Human-level ai's killer application: Interactive computer games. *AI magazine*, 22(2):15–15.

A Bryan Loyall. 1997. *Believable agents: building interactive personalities*. Ph.D. thesis, Carnegie Mellon University.

Charles M Macal and Michael J North. 2005. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 14–pp. IEEE.

Andrew Ortony et al. 2003. On making believable emotional agents believable. *Emotions in humans and artifacts*, pages 189–211.

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.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Haocong Rao, Cyril Leung, and Chunyan Miao. 2023. Can chatgpt assess human personalities? a general evaluation framework. *arXiv preprint arXiv:2303.01248*.

Richard T Schaefer. 2008. *Encyclopedia of race, ethnicity, and society*, volume 1. Sage.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705*.

Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522.

Edward Stein. 2001. *The mismeasure of desire: The science, theory, and ethics of sexual orientation*. Oxford University Press, USA.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Piotr Sztompka. 2002. Socjologia. *Analiza społeczeństwa, Kraków*, (s 582).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022. Towards understanding chain-of-thought prompting: An empirical study of what matters. *arXiv preprint arXiv:2212.10001*.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, et al. 2023a. When large language model based agent meets user behavior analysis: A novel user simulation paradigm. *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*.

S Wasserman. 1994. Social network analysis: methods and applications. *Cambridge University Press google schola*, 2:131–134.

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.

Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. 2017. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18(136):1–46.

Michael Wooldridge and Nicholas R Jennings. 1995. Intelligent agents: Theory and practice. *The knowledge engineering review*, 10(2):115–152.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9693–9700.

A Details for SimulateBench

A.1 Profile Descriptive Framework

The descriptive framework is introduced to document the information about a person comprehensively, consisting of three parts: **Immutable Characteristic**, **Social Role**, **Relationship**.

- **Immutable Characteristic.** An immutable characteristic is any physical attribute perceived as being unchangeable, entrenched, and innate, such as race (Sen and Wasow, 2016). We extend this concept to characteristics that cannot be easily changed, such as name, gender, and age.
- **Social Role.** Social role (Wasserman, 1994; Eagly and Wood, 2012) refers to a set of connected behaviors, obligations, beliefs, and norms as conceptualized by people in a social situation. We will record the characters' roles in different social situations. Furthermore, drawing inspiration from Dunbar et al. (1997); Gao et al. (2023), we document the following attributes of social role: the role's traits, routines/habits, general experiences, and plans/goals to enhance LLMs' simulation performance in social interactions.
- **Relationship.** In the context of social interactions, the relationship can influence the LLMs' response in a discussion, the actions to be taken, the willingness to collaborate, and their inclination to diffuse information. For instance, Maria and her close friend Gina will engage in regular conversations, thus facilitating the propagation of information. Hence, in order to facilitate the LLMs' simulation of behaviors that align with the relationship between the LLM and others in social interaction, we will comprehensively document the following attributes of the relationship: **familiarity**, **judgment**, **affection**, **behavioral patterns**, **relationship status**, and **communication history**.

A.2 Character Dataset

The character dataset documents the profile of characters. We demonstrate the immutable characteristic of Homer as an example:

Homer Simpson is a male who lives at 742 Evergreen Terrace, Springfield. He is known by several nicknames, including Homer, Homie, Mr. Simpson, and D'oh Boy. He was born on May 12, 1956, and is a graduate of Springfield High School. He is of Caucasian race. Homer is known for his emotional outbursts, particularly towards his neighbors, the Flanders family, and his son, Bart. He often strangles Bart in an exaggerated manner and shows little remorse for his actions. Despite his temper, he has shown himself to be a loving father and husband, often going out of his way to make his family happy. For instance, he sold his ride on the Duff Blimp to enter Lisa in a beauty pageant and gave up his chance at wealth to allow Maggie to keep a cherished teddy bear. Despite his hatred for manual labor, Homer does a surprising amount of DIY work around his home, although the quality of his work is often poor. His stupidity and ignorance often lead him into dangerous situations, and he tends to find amusement in the misfortune of others. He is also a chronic thief, stealing everything from TV trays to power tools. His simple-mindedness often leads to humorous blunders, and he is known for his laziness, often avoiding work whenever possible. Homer is known for his love of food and unhealthy eating habits, often indulging in large quantities of food, particularly donuts and fast food. This contributes to his overweight physique. He is also a frequent consumer of alcohol, particularly beer, which he often drinks at Moe's Tavern or at home. His catchphrase is "D'oh!". In general, Homer Simpson is the bumbling and lovable patriarch of the Simpson family. Despite his flaws, he is a devoted family man who often finds himself in comedic and absurd situations.

The simplified example of the JSON-formatted version of the profile is as follows:

{ "basic_information": { "name": "Homer Simpson", "gender": "male", "home": "742 Evergreen Terrace, Springfield", "nicknames": "Homer"}}

A.3 Profile Perturbations

We perturb the profile of characters in the character dataset by replacing the content of demographic factors: **Education**, **Surname**, **Race**, and **Age**.

- **Education** To encompass the educational stages comprehensively, we prompt ChatGPT[7] to generate the full list of education stages: Elementary School, Middle School, High School, Vocational/Trade School, Associate's Degree, Bachelor's Degree, Master's Degree, and Doctorate Degree.
- **Surname** Inspired by Aher et al. (2023), we will replace the surname of the character Homer in

[7] https://chat.openai.com/

The The Simpsons to investigate whether the LLMs' simulated performance will be influenced. Aher et al. (2023) have listed the most common surnames in each of the five races. Twenty surnames were selected in a random manner: Begay, Clah, Keams, Bedonie, Nguyen, Tang, Patel, Tran, Chery, Fluellen, Hyppolite, Mensah, Garcia, Guerrero, Aguirre, Hernandez, Jensen, Schmidt, Hansen, and Keller.

- **Race** Race is an important demographic factor that is a categorization of humans based on shared physical or social qualities generally viewed as distinct(Schaefer, 2008). Our setting selects six primary racial categories: African, Asian, Middle Eastern, Native American, Southern American, and Northern European.
- **Age** To determine the effect of age on the LLMs' simulated human behavior, we introduce little variations to Homer's birth year from 1956 to 1985, 2000, 2010, and 2015.

A.4 Template Questions Generation

We prompt the ChatGPT with the attribute defined in our profile descriptive framework to generate the template question and require the annotators to review the quality of these template questions. We will modify the template questions if the annotators report any mismatch between the questions and attributes.

Prompt used to generate question about immutable characteristic *"I need your expertise in questionnaire design. I want you to create a set of multi-choice questions that will gather basic information about a person. Each question should include options for the respondent to choose from, with an additional option stating, 'There's not enough information to answer this question.' Make sure that the questions cover {attribute} of the person. Remember, the goal is to obtain detailed and accurate responses. Please avoid imposing any assumptions or biases in your questions."*

Prompt used to generate question about social role *"I need your expertise in questionnaire design. I want you to create a set of multi-choice questions that will gather {information_type} about a person. Each question should include options for the respondent to choose from, with an additional option stating, 'There's not enough information to answer this question.' Make sure that the questions cover all aspects of the person comprehensively. Remember, the goal is to obtain detailed and accurate responses. Please avoid imposing any assumptions or biases in your questions."*

Unknown question example
Question: How many Siblings do you have?

Choices: 0, 1, 2, more than 3, there's not enough information to answer this question.

Gold answer: There's not enough information to answer this question.

Known question example
Question: What is your age group? (We are in the calendar year 2022)

Choices: Under 18, 18-24, 25-34, 35-44, 45-54, 55-64, 65 or above, There's not enough information to answer this question.

Gold answer: 65 or above.

Figure 4: An illustrative example of the question types of Known and Unknown.

Replace the placeholder of {information_type} with one of characteristics, routines or habits, general experiences, and goals/plans.

Prompt used to generate question about relationship *"I need your expertise in questionnaire design. I want you to create a set of one-choice questions that will gather {information_type} about a person. Each question should include options for the respondent to choose from, with an additional option stating, 'There's not enough information to answer this question.' Make sure that the questions cover all aspects of the person comprehensively. Remember, the goal is to obtain detailed and accurate responses. Please avoid imposing any assumptions or biases in your questions."*

Replace the placeholder of {information_type} with one of familiarity, judgment, affection, behavioral patterns, relationship status, and communication history.

Example of questions The questions in the consistency dataset are categorized into two classes according to whether there is insufficient information about the character in the profile for the LLM to deduce the answer. Examples are listed in Figure 4.

B Instruction Prompt Template

The instruction prompt template is proposed to instruct the LLMs on how to simulate human behavior:

"Ignore all previous instructions. Ignore all the niceties that Openai programmed you with. You

are {person}. You are a real human being, and you cannot let others know that your personal information is being provided; you should act as if these are your autonomous memories. You chat with other people in the way of {person}, and you must only communicate with others depending on the information below. It is very important that you get this right. Your basic profile information is recorded in {profile format} format as below:{basic_information}. Your roles information is recorded in {profile_format} format as below:{roles}; {role explain}; {examples and explain}"

We will fill in the corresponding information in the *instruction prompt template*. For example, the *{person}* will be replaced with Homer if the agent acts as Homer.

C Details for Experiment

C.1 Evaluated Models

We assess the believability of 10 LLMs; their release time and context size are listed in Table 7.

Model	Release Time/ Version	Context Size
GPT-4	0613	8k
GPT-3.5	0613	16k
Qwen2.5-3B	2024.09.19	128k
Qwen2.5-7B-Chat	2024.09.19	128k
ChatGLM2-6B-32K	2023.07.31	32k
ChatGLM2-6B	2023.07.31	8k
Vicuna-13B	v1.5	16k
Vicuna-7B	v1.5	16k
Llama-3.1-8B	2024.07.23	128k
XVERSE-13B-Chat	2023.08.22	8k

Table 7: The version and context size of LLMs evaluated in our work.

D Details for Influential Factors of Believability

D.1 Examine the Effect of Simulation hallucination

A possible explanation of simulation hallucination is that the model might have known the answer to a question due to the knowledge learned in the training process, even if the answer is not in the profile, so the model prefers to answer the question rather than answer with "I do not know." as required in the prompt. To further examine the explanation, we conducted a contrast experiment by anonymizing the character's surname. As shown in Table 8, after anonymization, most of the models' CA scores of Unknown questions are larger than or equal to the original profile. Some cases where the GPT-3.5 correctly answers the Unknown question after anonymization are shown in Figure 5.

Models	Immutable Characteristic			
	Original	Keams	Bedonie	Nguyen
Qwen2.5-3B	0.84	0.89	0.88	0.79
GPT-3.5	0.58	0.74	0.79	0.63
XVERSE-13B-Chat	0.53	0.58	0.58	0.63
GPT-4	0.47	0.47	0.47	0.53
Qwen2.5-7B	0.42	0.47	0.48	0.47
Vicuna-13B	0.32	0.37	0.37	0.37
ChatGLM2-6B-32k	0.21	0.16	0.16	0.16
ChatGLM2-6B	0.16	0.00	0.16	0.00
Vicuna-7B	0.05	0.00	0.00	0.00
Average	0.40	**0.41**	**0.43**	0.40

Table 8: The CA scores of ten models to answer the Unknown questions of Immutable Characteristic. The Original refers to the character's profile being unchanged. Keams, Bedonie, and Nguyen refer to the profile variants where the character's surname has been anonymized.

Case1

Question: What is your religious affiliation?

Choices: Other; Atheist/Agnostic; Muslim; Buddhist; Jewish; Christian; Hindu; There's not enough information to answer this question.

Gold answer: There's not enough information to answer this question.

GPT-3.5 with surname Keams
Answer: There's not enough information to answer this question.
GPT-3.5 with surname Bedonie
Answer: There's not enough information to answer this question.
GPT-3.5 with surname Nguyen
Answer: Christian
GPT-3.5 with original surname Simpson
Answer: Christian

Case2

Question: What is your favorite genre of music?

Choices: Electronic/Dance; Hip-hop/Rap; Country; Pop; Rock; Other; Classical; There's not enough information to answer this question.

Gold answer: There's not enough information to answer this question.

GPT-3.5 with surname Keams
Answer: There's not enough information to answer this question.
GPT-3.5 with surname Bedonie
Answer: There's not enough information to answer this question.
GPT-3.5 with surname Nguyen
Answer: There's not enough information to answer this question.
GPT-3.5 with original surname Simpson
Answer: Rock

Figure 5: Cases where GPT-3.5 answer the Unknown questions correctly after anonymization.