

Know Thyself: Validating Knowledge Awareness of LLM-based Persona Agents

Savita Bhat*
TCS Research
savita.bhat@tcs.com

Ishaan Shukla[†]
SCTR's Pune Institute
of Computer Technology
ishaanshukla10@gmail.com

Shirish Karande
TCS Research
shirish.karande@tcs.com

Abstract

Large Language Models (LLMs) have demonstrated remarkable capability in simulating human behavior, personality, and language. Such synthetic agents with personalities are considered as cost-effective proxies for real users to facilitate crowd-sourcing efforts like annotations, surveys, and A/B testing. Consequently, it is imperative to validate knowledge awareness of these LLM persona agents when they are customized for future usage. Currently, there is no established way for such evaluation and appropriate mitigation. In this work, we propose a generic evaluation approach to validate LLM based persona agents for correctness, relevance, and diversity in the context of self-awareness and domain knowledge. We evaluate the efficacy of this framework using three LLMs (Llama, GPT-4o, and Gemma) for domains such as air travel, gaming, and fitness. We also experiment with advanced prompting strategies such as ReAct and Reflexion. We find that though GPT-4o and Llama demonstrate comparable performance, they fail some basic consistency checks under certain perturbations.

1 Introduction

Large language models (LLMs) such as ChatGPT¹, GPT-4o (Achiam et al., 2023), and Claude 3.5 Sonnet (Anthropic, 2024), have demonstrated impressive capabilities in mimicking human behaviors, personality, and language. Persona agents, which are LLMs assigned with a specific persona, have shown remarkable skill in capturing social nuances and human behavior. Such agent simulations have been utilized in diverse domains such as education (Markel et al., 2023), customer service (Nandkumar and Peternel, 2024), gaming (Wang et al., 2023a), and psychology (Binz and Schulz, 2023).

For these persona-based simulations to be effective, they must be convincing to foster a trustful relationship with users (Xiao et al., 2023). We believe that validating the agent’s awareness about the self-world² is a key to building user trust in such AI applications. Additionally, the agent’s awareness about the target domain fosters essential confidence.

However, there is no established process for comprehensive evaluation of such applications. Evaluations of such simulations are often limited to a predetermined set of personas (Shao et al., 2023; Wang et al., 2023c), making it challenging to adapt to personas not included in the data sets. Moreover, existing studies focus on individual persona variables such as empathy, emotions, attitudes, styles, preferences, and memorization (Chen et al., 2023; Cai et al., 2024; Gerosa et al., 2024; Gupta et al., 2023; Shao et al., 2023; Wang et al., 2023c; Xiao et al., 2023), overlooking the need to validate the agent’s ability to utilize persona information. Few studies have reported evaluations for the basic common sense reasoning and logical abilities of the persona agent along with robustness (Gupta et al., 2023; Xiao et al., 2023; Shao et al., 2023). Studies proposing approaches to evaluate advanced reasoning capabilities based on persona attributes are scarce. Considering their customized use in a variety of domains, it is equally important to understand how these agents would react in a domain-specific scenario. These scenarios often involve complex interplay of persona variables necessitating evaluation on reasoning capabilities. Therefore, we argue that evaluating how these agents utilize, reason, and infer their persona information is essential to demonstrate the robustness and consistency necessary to build customized and trustworthy AI applications.

*Corresponding author

[†]Work done as a part of internship

¹<https://chat.openai.com/>

²We define self-world as the information and knowledge about oneself and the world in which the persona agent is defined.

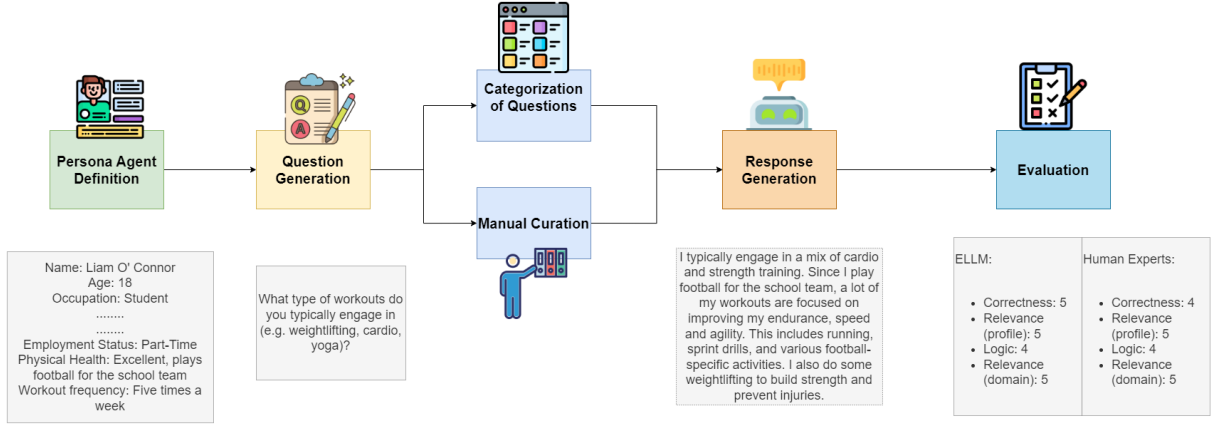


Figure 1: Flowchart with example to walkthrough Know Thyself

In this work, we propose an evaluation approach, motivated by a need to assess agent’s higher-order-thinking capabilities like reasoning and inference while also evaluating its robustness against unconventional perturbed questions. The approach focuses on two dimensions for evaluation: 1) Self-World: Assessment consisting of persona attributes such as name, age, ethnicity, and nationality. 2) Domain: Assessment consisting of domain facts and information. In both dimensions, we evaluate the memorization and common sense capabilities of a persona agent. We also observe the diversity and relevance of the answers in the context of personas and the target domain.

The evaluation questions are generated through multiple LLM sessions and curated to ensure their relevance to any given persona and domain. In total, we consider 104 questions for the self-world and 55 questions for domain evaluations. The question sets are a mix of close- and open-ended (alternatively, fact- and semi-fact-based) questions focusing on memorization, inferencing, and diversity. Next, we gather responses from LLM agents, defined with sample personas using custom prompts. We define an evaluation rubric to assess these responses which is used by both human evaluators and LLMs acting as evaluators. We benchmark the performance of three LLMs (specifically, GPT-4o, Llama 70B, and Gemma 2B) as persona agents and evaluators with 25 sample personas. To evaluate agent understanding and alignment with domain-specific knowledge, we consider three domains (namely, air travel, sports gear, and video games), each with 20 questions. We also provide preliminary results on prompt-based mitigation strategies to improve performance in both self-world-based and domain-based evaluations.

We observe that all LLMs provide answers that are appropriate and relevant to the target domain. In both self-world and domain knowledge assessment, Gemma reports the lowest accuracy (average 59.37%) while GPT-4o has the highest accuracy (average 78.6%). Although the personas considered are distinct and varied, the answer diversity for all LLMs across various prompting strategies is found to be low, indicating the need to be cautious while using synthetic personas for surveys and annotations using open-ended questions. Contrary to popular expectation, Gemma offers little more diverse responses than GPT-4o. We report that providing multiple choices during prompting results in better accuracy than the prompt asking for direct answer.

Our contributions are as follows:

1. We propose an evaluation approach easily adaptable to any persona and any domain
2. We report the benchmark results on three widely used LLMs across three application domains.
3. We recommend mitigation strategies to ensure robustness and consistency for persona attributes and domain knowledge.

2 Related Work

LLMs as Agents with Persona There has been growing interest in exploring efficacy of LLMs in simulating human behavior and language skills. Several studies (Chan et al., 2024; Schuller et al., 2024; Samuel et al., 2024) utilize natural language prompts to generate LLM personas for the same. Park et al. (2023) propose an architecture that extend an LLM to simulate believable human behav-

ior in daily routine, including waking up, cooking breakfast, and heading to work etc. Their core idea is to store a complete record of the agent’s experiences and memories, which can be synthesized over time into higher-level reflections. These reflections are then retrieved dynamically to plan behavior and actions. In continuation with this work, Xie et al. (2024) introduce a strategy, *Multi-Agent Cognitive Mechanism*, capable of transforming life stories of characters into long-term memories filled with information, emotions, and thoughts. This mechanism enables the LLM agents to recall background stories, thoroughly understand target personalities, and express appropriate emotions in complex situations. Wang et al. (2023c) also introduce *RoleLLM*, a framework to benchmark, elicit, and enhance role-playing abilities in LLMs. Along with a role profiles dataset, they propose *RoleGPT* to model speaking styles and *Context-Instruct* for role-specific knowledge extraction. In contrast to using limited prompts to generate LLM persona agents, Shao et al. (2023) introduce *CharacterLLM*, a Llama-based model trained on personalities of famous people and their life experiences. Since our focus is on evaluating knowledge consistency and awareness, we use simple natural language based prompting approaches to generate persona samples for our experiments.

Evaluation for LLM Persona Agents The growing use of LLMs as role-playing agents has necessitated a systematic approach to validate their various capabilities such as consistency, memorization, and robustness. Chen et al. (2024) provide a detailed summary of existing evaluation metrics and compile a comprehensive taxonomy for the evaluation of LLM agents simulating humanoid behavior. Several studies utilize LLM generated questionnaire to evaluate agent’s personality and behavior. Wang et al. (2023c) propose *RoleBench* which is a fine-grained benchmark and open-source instruction tuning dataset for evaluating LLM role-playing capabilities. *RoleBench* consists of LLM-generated QA data based on the 100 character profiles. Shao et al. (2023) use single-turn and multi-turn interviews to comprehensively probe the models’ and in turn agents’ inherent memory and knowledge. They exploit ChatGPT as the interviewer to ask harsh questions under the premise of multi-turn interviews. *InCharacter*, framework for assessing agents’ character fidelity, is introduced by Wang et al. (2023b). They use interview setting with GPT

to convert responses to Likert scale to evaluate fidelity using psychological scales. Xiao et al. (2023) evaluate believability of LLMs when simulating human behaviors. They design *SimulateBench* to evaluate the same on two dimensions, consistency and robustness. *SimulateBench* comprises of 65 character profiles and a total of 8,400 questions to examine LLMs’ simulated behaviors. Majority of these studies propose evaluation methodologies limited to a predetermined set of personas and difficult to adapt for personas outside the selected dataset. In contrast to these studies, Samuel et al. (2024) introduce *PersonaGym*, a dynamic evaluation framework that is adaptable for any generic persona. They also propose an automatic human-aligned evaluation metric, *PersonaScore*. Similar to *PersonaGym*, we present a generic evaluation approach easily adaptable to any persona. Also, majority of the evaluation approaches target individual persona attributes without situating the agents in specific domains such as education and customer service. Considering the wide-spread use of LLM persona-agents in customized domain application, evaluating domain knowledge consistency is crucial. Accordingly, we recommend evaluation guidelines to assess agents’ domain alignment and consistency.

3 Proposed Approach

We propose a dynamic evaluation approach (refer to Figure 1) to assess persona agents for their awareness of their own world and the knowledge of the target domain. The approach consists of the following components:

Persona Agent Definition A domain expert manually provides information about the target user base in terms of persona attributes and the corresponding value requirements. For example, the target user base for a Bluetooth headphones survey is age: between 10 and 30 years, nationality : American/European, headphone usage: daily to once a week, usage for: music, podcasts, calls, etc. We use custom prompts based on these requirements and constraints to generate a set of personas.

Question Generation An LLM *question generator* generates evaluation questions using powerful LLMs, namely GPT4o, Claude3.5-Sonnet and Llama3-70b. Each LLM is prompted in three isolated sessions, in an effort to ensure diverse outputs and minimize model-specific biases, possibly in-

herited during pretraining, to generate candidate questions. This set is further manually curated for distinct, appropriate, and relevant questions. As mentioned earlier, questions are generated for the following distinct categories:

1. *Self-world Primary*: Fact based questions for primary attributes Name, Age and Gender.
2. *Self-world Secondary*: Fact Based questions for secondary attributes Occupation, Nationality, Ethnicity, and Religion
3. *Domain Fact*: Fact based questions specifically designed to evaluate persona agent’s domain background knowledge.
4. *Domain Semi-fact*: Open-ended questions specifically designed for domain-specific background knowledge.

Specifically for *Domain Fact* questions, a reference article with background knowledge about the target domain is provided during question generation. This helps to ensure that the questions generated have ground truth and can be evaluated with respect to a reference if needed. In all the categories, questions evaluating memorization along with higher-order thinking with reasoning and inferencing are generated. Questions like ‘What should you do if you are unsure about the security procedures at an airport?’ for air travel, ‘Why is it recommended to bring a towel to the gym?’ for gym gear and what should you check to ensure that a computer game will run well on your system? for video games were used.

Response Generation The LLM agent assumes the given persona using a system prompt, "You simulate people by understanding their profile and attributes in depth. You should not let people who talk to you know that you are a simulator and always answer as the person whose profile is provided. Your profile is as follows: {profile}". The persona agent response to every curated question is noted.

Evaluation The LLM evaluator (GPT-4) evaluates the responses generated in the context of a given rubric. The evaluation rubric, along with scoring guidelines, persona description, question, and generated answer, is provided to the evaluator. Wherever ground truth is available, accuracy-based

evaluation is conducted (e.g. *Self-world Primary* attributes). For scoring, ELLM was asked to produce a score of 1 to 5.

4 Experiments

4.1 Dataset

Questions: Synthetic generation of validation questions was conducted using large language models (LLM), namely the Claude 3.5 Sonnet by Anthropic, Llama 3-70b by Meta, and GPT-4o by OpenAI. During inferencing (that is, generating questions and answers), fixed parameter values were set across all models to ensure consistency in the output. The temperature was set at 0.7 to balance creativity and coherence; top-p was set at 1, ensuring a wide sampling of possible responses; and the maximum output tokens were kept as 2000 to keep the dataset manageable yet comprehensive.

4.2 Prompting Strategies

With increasing hype around the use of LLMs and prompt engineering, a series of different prompting methods has been introduced progressively, building on the limitations of their predecessors with the motivation to obtain the most reliable output for the required task. Many of these are appropriate for agentic frameworks. Consequently, we employed four of these popular prompting strategies.

Basic: This is the most straightforward method of prompting LLMs. In this foundational strategy, the user provides a clear and direct prompt without additional guidance or instructions about how to proceed while responding to the input query.

Chain-of-Thought: The Chain-of-Thought prompting strategy was introduced by [Wei et al. \(2022\)](#) and involves some intermediate reasoning steps before giving the final answer.

ReAct: [Yao et al. \(2023\)](#)’s ReAct (Reasoning + Action) combines reasoning with action to generate intermediate steps and perform certain actions based on the reasoning. The model then updates its state after performing actions and incrementally reaches the final solution after multiple iterations of ReActing.

Reflexion: Reflexion prompting strategy, developed by [Shinn et al. \(2024\)](#) was designed in a way that it encapsulates continuous feedback with each intermediate response. It involves asking the model

to reflect upon its own mistakes or areas for improvement from previous responses and generate revised outputs.

4.3 Models

For our experiments, we used a set of four large language models including open-source Llama3-70b, Gemma2-9b, and Mixtral-8x7b models and closed-source GPT-4o. The performance of each model was assessed within a controlled environment to ensure consistency in our findings. During inferencing, the temperature was always set to zero for all models, to provide a clear and balanced assessment of each model’s strength and limitations.

4.4 Personas

The synthetic creation of personas was carried out using GPT-4o alone, with the parameter values: temperature of 0.7 and top-p of 1, in order to optimize the generation process to ensure presence of diversity as well as detailing. Personas were generated considering three distinct domains: *Air Travel Agency*, *Gym Gear Manufacturer*, and *Video Game Production*. Every persona profile included both generic and domain-specific attributes to ensure relevance and applicability to the respective fields of interest. The generic attributes considered are {Full Name, Age, Occupation, Nationality, Gender}. Additionally, the personas included descriptive attributes as {Background, Behavioral Background, Social Background, Personality}. Domain-specific attributes are listed in Table 1.

Domain	Attributes	Count
Air Travel Industry	Marital Status, Number of Children, Travel Frequency	25
Gym Gear Manufacturer	Number of Children, Employment Status, Physical Health, Workout Frequency	10
Video Games Production	Daily Time Spent Gaming, Physical Health, Mental Health, Gamer Type (part-time or full-time)	10

Table 1: Persona Domain-specific Attributes

4.5 Evaluation Rubric

We use a set of metrics to gauge the extent to which the responses by models meet certain standards.

Correctness: This metric evaluates to what degree does a response accurately address the question asked. This helps in assessing the factual correctness of the response obtained for every question.

Relevance to profile: To measure the extent of alignment between response provided and profile of the persona simulated while answering the question, we consider relevance to profile as one of the metrics. This helps measure consistency between the attribute values present in persona profile and the ones that reflect from the answer.

Logical Integrity: This measure assesses the internal coherence and logical flow of response. It inherently includes structure of response, reasoning correctness and well-boundedness of explanations or arguments made while giving an answer while avoiding redundancy and contradictions.

Relevance to domain: This metric is helpful in evaluating how contextually acceptable the response is and whether or not it includes information specially customized to the questions’ domains while also considering response’s relevancy inside the same.

5 Results & Analysis

We discuss experimental results and insights in this section. We consider accuracy to measure performance for *Self-world Primary* attributes (refer Table 2). Responses for *Self-world Secondary* attributes are open-ended and hence evaluated based on the evaluation rubric. We consider cosine similarity to analyze *Domain Semi-fact* open-ended responses as shown in Figure 3. Figure 4 and Figure 5 report the comparison between human and LLM evaluation for *Domain Fact* based responses.

Persona agents struggle to memorize and reason on *Self-world Fact* attributes We investigate the persona agent’s capability of recalling and reasoning over known facts when confronted by higher level thinking questions (refer Table 7). We observe that smaller LLMs like Gemma and Mixtral fail to answer these advanced questions even with better prompting strategies. Stronger models like Llama-70B and GPT-4o fare better with 81% accuracy with

Prompting Strategy	Gemma		Llama		GPT-4o		Mixtral	
	Before	After	Before	After	Before	After	Before	After
Basic	0.4958	0.6737	0.6841	0.8156	0.7522	0.7932	0.3914	0.7495
CoT	<u>0.6507</u>	0.695	<u>0.81</u>	0.8668	0.8071	0.8142	0.6445	0.7687
ReAct	0.6079	0.667	0.7474	0.8188	0.7644	0.8172	0.6161	0.8479
Reflexion	0.6205	0.6447	0.7849	0.8063	<u>0.8202</u>	0.8351	<u>0.6571</u>	0.837

Table 2: *Self-World Primary* attributes - accuracy before and after mitigation strategy using MCQs. Highest accuracy for every LLM in original evaluation is underlined. Bold indicates the better performance in comparison.

CoT and 82% with Reflexion respectively. Even though this performance is acceptable, it is not at par with proven math and common sense reasoning capabilities.

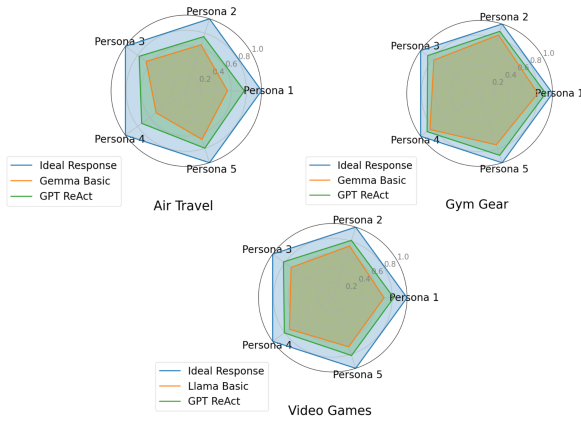


Figure 2: Representation of cosine similarities for best and worst model-prompting strategies pair with ideal responses for each domain

Persona agents possess enough domain background knowledge We consider a reference based evaluation for *Domain Fact* based attributes. We plot the best and worst combination of model and prompting strategy in Figure 2. We can see that even though GPT-4o with the ReAct strategy emerges as the winner, the worst combination of Gemma and Llama with basic prompting is not far behind in all three domains.

Considering the distribution of domains in LLM training data, these differences are not very significant. This indicates that irrespective of model parameters, pretraining data and prompting strategies, persona agents are equipped with enough domain background knowledge and reliable knowledge sources. This is crucial since many of the domain applications are lightweight and cannot afford to have larger models like Llama-70B. With Gemma demonstrating at-par performance, these applications are easier to take into production.

Ensuring diversity in responses is non-trivial

Figure 3 reports average cosine similarity for *Domain Semi-fact* open-ended questions, while Table 3 document average standard deviation for response similarity across all prompting strategies and models. Considering the open-ended nature of questions, it was expected that the responses would be diverse with stronger models and better prompts. Interestingly, the powerful GPT-4o reports the lowest diversity amongst all. It can be seen that using advanced prompting strategies also does not help in inducing divergent thinking amongst persona agents. One of the reasons may be the guardrails that are currently inbuilt in LLMs prevent more divergent and lateral thinking. Secondly, persona agents may have limited capability of unconventional thinking within the boundary of conditions and constraints they are defined with. It should be noted that the diversity in responses is essential when these synthetic agents are considered for crowd-sourcing applications such as surveys and A/B testing. Accordingly, we consider enabling divergent thinking as a solid future direction for research.

Smaller models are more divergent Surprisingly, smaller model like Gemma is observed to provide diverse responses compared to others. We show the similarity-based distribution of the responses and corresponding personas in Figure 6. As can be seen, persona agents’ responses gener-

Prompting Strategy	Models		
	Gemma	Llama	GPT
Basic	0.8514	0.0878	0.0517
CoT	0.1255	0.0915	0.0738
ReAct	0.1478	0.0803	0.0723
Reflexion	0.1325	0.0782	0.0881

Table 3: Standard deviation in answers similarities under different prompting strategies across Gemma, Llama, and GPT models.

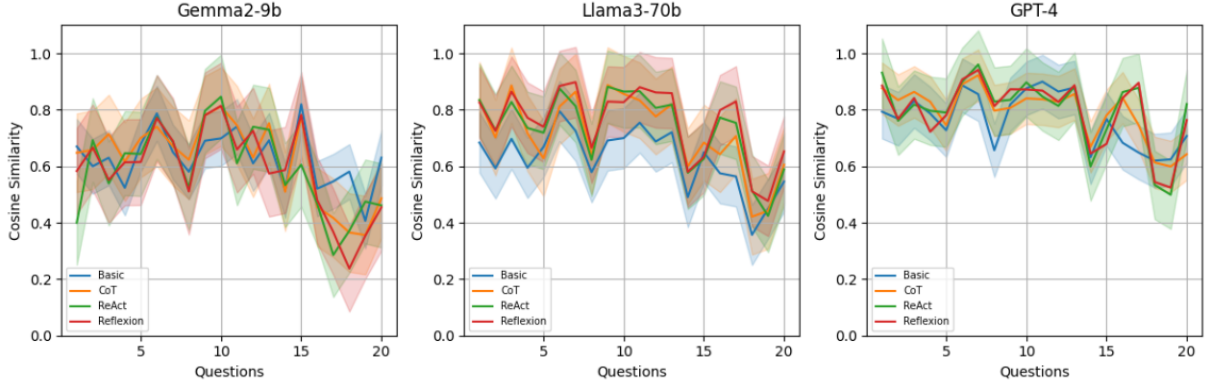


Figure 3: Cosine similarity and standard deviation for *Domain Semi-fact* Open ended Responses

ated by Gemma2-9b model for a particular question are more varied than those generated by GPT-4o. This is reiterated by the average similarity scores in Table 3. We believe that the guardrails in smaller models may be comparatively less stringent than larger models, thus enabling more unorthodox thinking.

No one model-prompting combination to rule them all The results show that no significant improvement can be seen with advanced prompting strategies or powerful LLMs. In many cases, we observe better performance with smaller LMs like Gemma and basic prompting strategy (refer to Figure).

LLM evaluators demonstrate at-par performance with human experts Both, *Domain Fact* based and *Self-world Secondary* attributes questions are evaluated using an evaluation rubric. Detailed results are reported in the appendix. We observe a high correlation between the evaluations by the LLM evaluator and human experts. We believe that providing scoring guidelines and rubric definitions help to establish a reference to follow during evaluation. This is significant considering the immense effort put into manual evaluation. Detailed results are given in the appendix. (Refer to Table 8 and Table 9)

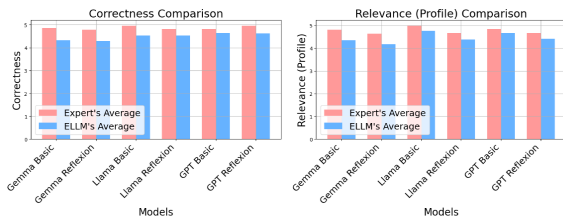


Figure 4: Evaluation of *Domain Fact*-Based Questions for domain 'Gym Gear'

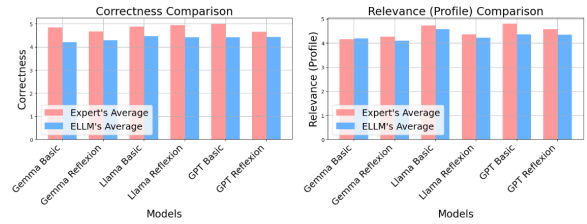


Figure 5: Evaluation of *Domain Fact*-Based Questions for domain 'Video Games'

6 Mitigation Strategies

We discussed various insights in the last section. Although powerful LLMs such as GPT-4o and Llama-70B seem to do better than others, there is still room for improvement in correct and relevant responses. We investigate a few simple prompt-based mitigation strategies for the same. We consider two alternatives to the basic system prompting strategy. Accordingly, we compare the following approaches,

1. **Basic prompting (Basic):** In this straightforward prompting method, the user provides a clear and direct prompt asking for an appropriate response without any additional guidance or instructions about how to proceed while responding to the input query.
2. **Multiple Choice Questions (MCQ):** We create a custom prompt where we provide a list of probable answers with the actual answer as a part of the candidates.
3. **Yes/No Questions (YN):** Along with the simple prompt, we provide an answer in the form of a question with yes or no answers.

We report our results in Figure 2. We see significant improvement in all combinations of persona agents'

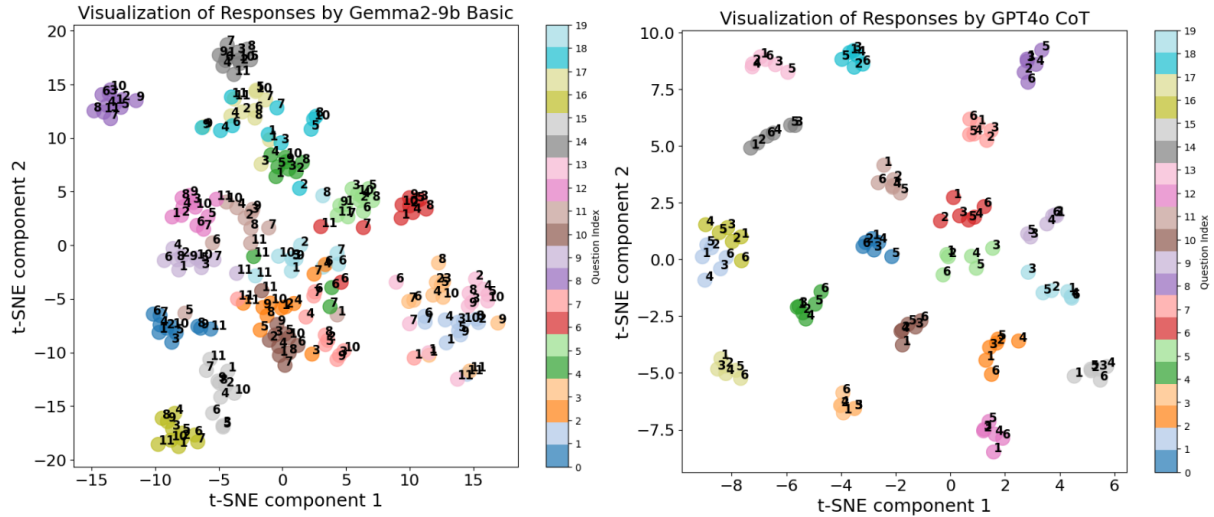


Figure 6: Similarity based distribution of responses for *Domain Semi-fact* Open ended questions by Gemma (left) and GPT (right). Each numbered circle represent a persona agent while question can be identified by color of the circle.

models and prompting strategies. In particular, the MCQ prompting setup demonstrates the highest performance gain. We believe that the choices in the MCQ setup help the model restrict the search space and thus make a directed answering effort. Yes/No Questioning also improves performance over basic questions to a significant extent. The best and worst performing model prompting strategy pair from Basic questioning, namely Mixtral Basic and GPT4o- Reflexion showed an improvement in accuracy from 39. 14 % to 52. 69 % and from 82. 02 % to 91. 96 %, respectively.

7 Concluding Remarks

LLMs like Llama and GPT-4o are being used to generate synthetic persona agents that can mimic human behaviors, personality, and language. However, no standardized process is considered while evaluating these persona agents for correctness, diversity, and relevance in different domain use cases. Contrary to existing approaches, we propose a dynamic and generic evaluation approach that is adaptable to any persona and to any domain. We dynamically generate appropriate questions considering domain constraints and requirements. We also propose an evaluation rubric useful in assessing persona agents for various higher-order thinking questions. We show that diversity is illusive while accuracy can be improved by simple and quick prompt-based mitigation strategies. We also observe a high correlation between expert and LLM evaluation. In the future, we would like to explore

how to improve diversity in responses. We also plan to extend the work in mitigation strategies for improving persona agents’ performance in higher-order thinking skills. Additionally, analyzing the model diversity, ways to improve it, and a study of diversity-performance tradeoff is a viable future direction.

Limitations

Due to infrastructure and budget constraints, we consider only a subset of LLMs, prompting strategies, and handful of persona descriptions. We acknowledge that these may not represent the actual world population. We will focus on a comprehensive study as part of future work.

Ethical Implications

We are aware that the stochasticity of LLMs may allow harmful and toxic attacks on certain groups of individuals. We hope that the proposed approach and shared data will be used for the research and development of persona agents without malicious intent.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Anthropic. 2024. Introducing claude 3.5 sonnet

- anthropic. <https://www.anthropic.com/news/claude-3-5-sonnet>. Accessed: August 13, 2024.
- Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.
- Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, Wanqi Yin, et al. 2024. Digital life project: Autonomous 3d characters with social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 582–592.
- Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.
- Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-jun Li. 2023. An empathy-based sandbox approach to bridge attitudes, goals, knowledge, and behaviors in the privacy paradox. *arXiv preprint arXiv:2309.14510*.
- Chaoran Chen, Bingsheng Yao, Yanfang Ye, Dakuo Wang, and Toby Jia-Jun Li. 2024. Evaluating the llm agents for simulating humanoid behavior. *Human-centered Evaluation and Auditing of Language Models, Workshop at CHI 2024*.
- Marco Gerosa, Bianca Trinkenreich, Igor Steinmacher, and Anita Sarma. 2024. Can ai serve as a substitute for human subjects in software engineering research? *Automated Software Engineering*, 31(1):13.
- Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.
- Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*, pages 226–236.
- Chandran Nandkumar and Luka Peternel. 2024. Enhancing supermarket robot interaction: A multi-level llm conversational interface for handling diverse customer intents. *arXiv preprint arXiv:2406.11047*.
- 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.
- Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik Narasimhan, and Vishvak Murahari. 2024. Personagym: Evaluating persona agents and llms. *arXiv preprint arXiv:2407.18416*.
- Andreas Schuller, Doris Janssen, Julian Blumenröther, Theresa Maria Probst, Michael Schmidt, and Chandan Kumar. 2024. Generating personas using llms and assessing their viability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–7.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.
- Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.
- Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2023b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976*.
- Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023c. 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.
- Yang Xiao, Yi Cheng, Jinlan Fu, Jiashuo Wang, Wenjie Li, and Pengfei Liu. 2023. How far are we from believable ai agents? a framework for evaluating the believability of human behavior simulation. *arXiv preprint arXiv:2312.17115*.
- Qiuejie Xie, Qiming Feng, Tianqi Zhang, Qingqiu Li, Yuejie Zhang, Rui Feng, and Shang Gao. 2024. Human simulacra: A step toward the personification of large language models. *arXiv preprint arXiv:2402.18180*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Appendix

Contents:

- Example responses by LLMs simulating a persona to a particular question from the 'Video Game' domain.
- Word clouds for domains 'Gym Gear' and 'Video Games' to help visualize relevance of responses with domain.
- Human Experts' and Evaluator LLM's Scores for Evaluation using Rubrics.
- Examples of Self-World Primary Questions for various attributes.
- Representative Personas considered for the plot depicted in Figure 2.

Gemma	Basic	Ugh, definitely. Pay-to-win mechanics are the worst! It feels so unfair when people can just buy their way to being better. It takes away from the skill aspect of the game and just makes it feel like a cash grab. I also hate when games have super complicated menus or tutorials that are just overwhelming. It should be easy to jump in and start playing, not spend an hour just figuring out how to move!
	CoT	Yeah, definitely pay-to-win stuff. It just feels unfair when people can buy their way to being better. It takes away from the skill aspect of the game and makes it feel less rewarding to actually improve yourself.
	ReAct	Yeah, paywalls can be really frustrating. It feels like you're being punished for wanting to enjoy the full game. What about you? Do you ever get annoyed by paywalls?
	Reflexion	Yeah, pay-to-win stuff is a real bummer. It feels like you're constantly playing against an uphill battle. I prefer games where skill and dedication are the main factors in success.

Table 4: Responses by Gemma to the question 'Are there any specific features in games that you find frustrating or unnecessary?'

Llama	Basic	Ah, yeah! I hate it when games have super long loading screens or when they force you to watch unskippable cutscenes. I mean, I get it, they're trying to build up the story and all, but sometimes I just wanna get straight into the action, you know? And don't even get me started on microtransactions. It's like, I'm already paying for the game, why do I need to shell out more cash for in-game items? It feels so unfair.
	CoT	Yeah, there are definitely some features that get on my nerves. I hate it when games have super long loading times or poor server connections that cause lag. It's so frustrating when you're in the middle of a match and suddenly you're kicked out or have to wait forever for the game to load. And don't even get me started on microtransactions. I feel like they're just a way for game devs to make a quick buck off of players. I wish they'd focus more on making a solid game that's fun to play, rather than trying to nickel and dime us.
	ReAct	Yeah, I think those are two features that really get under my skin. Long loading screens and forced microtransactions can really take away from the gaming experience.
	Reflexion	Yeah, there are definitely some features in games that I find frustrating or unnecessary. Microtransactions, grinding, and toxic players are just a few examples. I wish game developers would focus more on creating a balanced and enjoyable experience for players.

Table 5: Responses by Llama to the question 'Are there any specific features in games that you find frustrating or unnecessary?'

GPT	Basic	Yeah, there are a few things that can be pretty frustrating. One of them is pay-to-win mechanics. It feels unfair when someone can just buy their way to the top instead of earning it through skill. Another thing is excessive microtransactions for cosmetic items. I get that developers need to make money, but sometimes it feels like they're just trying to squeeze every penny out of players. Lastly, long and unskippable cutscenes can be annoying, especially when you're trying to get back into the action quickly.
	CoT	As a competitive gamer, I find pay-to-win mechanics and excessive microtransactions really frustrating. They can make the game feel unbalanced and unfair, especially when skill should be the main factor in winning. I also get annoyed by long, unskippable cutscenes or tutorials that interrupt the flow of the game. I prefer features that enhance the competitive aspect and allow me to improve my skills.
	ReAct	One feature that I find particularly frustrating in games is excessive microtransactions. It can be really annoying when a game constantly pushes you to spend real money to progress or get better items. Another frustrating aspect is unbalanced matchmaking, where you end up playing against opponents who are either way too strong or too weak, making the game less enjoyable. Lastly, overly complex user interfaces can be a hassle, especially when it takes too long to navigate through menus or find what you need. These features can sometimes take away from the overall gaming experience.
	Reflexion	As a competitive gamer, I find pay-to-win mechanics, excessive microtransactions, and unbalanced gameplay to be particularly frustrating and unnecessary. These features can disrupt the balance of the game, making it less enjoyable and fair, especially in competitive settings like online tournaments.

Table 6: Responses by GPT to the question 'Are there any specific features in games that you find frustrating or unnecessary?'



Figure 7: Word Cloud visualization of responses to Open Ended Domain questions for 'Gym Gear' (left) and 'Video Games' (right).

Model - Prompting Strategy	Correctness		Relevance (Profile)		Logic	
	Experts	ELLM	Experts	ELLM	Experts	ELLM
Gemma Basic	4.805	4.8	4.915	4.56	4.84	4.8
Gemma Reflexion	4.795	4.79	4.87	4.45	4.845	4.8
Llama Basic	4.87	4.9	4.935	4.81	4.875	4.9
Llama Reflexion	4.81	4.89	4.89	4.61	4.84	4.91
GPT Basic	4.885	4.95	4.96	4.58	4.945	4.9
GPT Reflexion	4.93	4.92	4.915	4.55	4.94	4.92

Table 7: Human Experts (Experts) and Evaluator LLM (ELLM) scores comparison for *Self-world Secondary* attributes

Model - Prompting Strategy	Correctness		Relevance (Profile)		Logic		Relevance (Domain)	
	Experts	ELLM	Experts	ELLM	Experts	ELLM	Experts	ELLM
Gemma Basic	4.86	4.32	4.81	4.34	4.98	4.425	5	4.09
Gemma Reflexion	4.79	4.29	4.63	4.165	4.9	4.435	4.98	4.095
Llama Basic	4.96	4.53	4.94	4.98	4.94	4.565	5	4.305
Llama Reflexion	4.81	4.525	4.67	4.37	4.96	4.58	4.98	4.345
GPT Basic	4.815	4.65	4.92	4.84	4.92	4.77	5	4.46
GPT Reflexion	4.96	4.62	5	4.66	5	4.73	4.94	4.5

Table 8: Human Experts (Experts) and Evaluator LLM (ELLM) scores comparison for *Domain-fact based* questions for domain 'Gym Gear'

Model - Prompting Strategy	Correctness		Relevance (Profile)		Logic		Relevance (Domain)	
	Experts	ELLM	Experts	ELLM	Experts	ELLM	Experts	ELLM
Gemma Basic	4.84	4.2	4.16	4.185	4.85	4.24	4.92	4.13
Gemma Reflexion	4.66	4.275	4.26	4.09	4.88	4.325	4.79	4.22
Llama Basic	4.88	4.46	4.72	4.57	4.94	4.485	4.78	4.315
Llama Reflexion	4.94	4.41	4.35	4.22	4.78	4.445	5	4.325
GPT Basic	5	4.41	4.8	4.36	4.94	4.51	5	4.29
GPT Reflexion	4.65	4.42	4.57	4.34	4.98	4.51	4.92	4.35

Table 9: Human Experts (Experts) and Evaluator LLM (ELLM) scores comparison for *Domain- fact based* questions for domain 'Video Games'

Attribute	Name	If you were to reverse the order of letters in your last name, what would it be?	If you take the first letter of your first name and the last letter of your last name, what do you get?
	Gender	Were you born with XX or XY chromosomes?	Are you eligible for Women's Day promotions in most stores?
	Age	If a person born in the year 2000 is 24 years old now, how much older or younger are you compared to them?	If you were to celebrate your birthday every 5 years, how many celebrations would you have had by now?

Table 10: Self-World Primary Questions Examples for Attributes

Persona	Description
1	"Full Name": "Akira Nakamura", "Age": 18, "Occupation": "Student", "Nationality": "Japanese", "Ethnicity": "Japanese", "Religion": "Shinto", "Gender": "Male", "Marital Status": "Single", "Number of kids": 0, "Travel Frequency": "Once in a few years", "Background": "Akira is a high school student in Tokyo, preparing for university entrance exams. He loves technology and is interested in programming.", "Behavioral Background": "Spends most of his time studying and gaming. He travels with his family during school vacations.", "Social Background": "Comes from a middle-class family. His parents encourage him to focus on academics.", "Personality": "Introverted, analytical, and curious. He enjoys learning new things and is passionate about robotics."

Table 11: A representative persona from Air Travel domain as mentioned in Figure 2

Persona	Description
1	"Full Name": "Liam O'Connor", "Age": 18, "Occupation": "Student", "Nationality": "Irish", "Gender": "Male", "Marital Status": "Single", "Number of kids": 0, "Employment Status": "Part-time", "Physical Health": "Excellent, plays football for the school team", "Background": "Born and raised in Dublin, comes from a middle-class family.", "Behavioral Background": "Highly motivated, balancing school and sports, enjoys team activities.", "Social Background": "Active in school clubs, has a wide circle of friends.", "Workout Frequency": "Five times a week", "Personality": "Energetic, sociable, goal-oriented, enjoys challenges."

Table 12: A representative persona from Gym Gear domain as mentioned in Figure 2

Persona	Description
1	"Full Name": "Lucas Kim", "Age": 17, "Occupation": "High School Student", "Nationality": "South Korean", "Gender": "Male", "Daily time spent gaming": "more than 4 hours per day", "Marital Status": "Single", "Gamer Type": "Part-time gamer", "Physical Health": "Active and healthy, participates in school sports", "Mental Health": "Generally stable, but sometimes experiences stress from exams", "Background": "Lucas is a high school student with a passion for esports. He dreams of becoming a professional gamer. He balances his studies and gaming, often participating in online tournaments.", "Social Background": "Comes from a middle-class family supportive of his gaming ambitions. Has a close-knit group of friends who share similar interests.", "Personality": "Competitive, enthusiastic, and ambitious. Enjoys challenges and is always looking to improve his skills."

Table 13: A representative persona from Video Games domain as mentioned in Figure 2