Examining Persona Drift in Conversations of LLM Agents

Anonymous ACL submission

Abstract

Large Language Models (LLMs) show impressive conversational abilities but sometimes show 'persona drift' problems, where their interaction patterns or styles become inconsistent over time. As the problem has not been thoroughly examined yet, this study examines consistency of expressed persona across nine LLMs. Specifically, we (1) investigate whether LLMs could maintain consistent patterns in expressed persona and (2) analyze the effect of the model family, parameter sizes, and types of given persona. Our experiments involve multi-turn conversations on personal themes, analyzed in qualitative and quantitative ways. Experimental results indicate three findings. (1) Larger models experience greater persona drift. (2) Model differences exist, but their effect is not stronger than parameter sizes. (3) Assigning a persona may not help to maintain persona expressions. We hope these three findings can help to improve persona consistency in AI-driven dialogue systems, particularly in long-term conversations.

1 Introduction

Recent research has actively explored the utilization of Large Language Models (LLMs) as chatbot systems by assigning them specific personas (Samuel et al., 2024; Nandkumar and Peternel, 2024; Tseng et al., 2024). To enhance user satisfaction in such systems, maintaining the consistency of the persona given to the LLM is critical. If the persona of an LLM loses its consistency, it may fail to deliver the user experience expected by the users, leading to usability issues (Tanprasert et al., 2024). So, researchers recently focused on investigating whether LLMs can preserve persona during a conversation, focusing on two aspects of persona (Tseng et al., 2024; Chen et al., 2023; Maharana et al., 2024; Zhang et al., 2024b; Afzoon et al., 2024): (1) memory, which is inputted to LLMs to maintain content consistency in a conversation, and (2) expressed persona, which is related to behavioral or stylistic consistency in a conversation and can be observed from output of LLMs. Among the two aspects, we focus on whether LLMs can retain the expressed persona.

Regarding the expressed persona, existing studies focused on LLMs' persona (Huang et al., 2023; Wang et al., 2024; Zhang et al., 2024a; Frisch and Giulianelli, 2024) without any conversation. Mainly, most researchers examined which persona LLMs exhibit in a specific isolated situation. Though existing work revealed LLMs have a consistent persona without any interaction, it is questionable whether LLMs can retain such persona expression throughout a long conversation. As many reports suggest that LLMs are very sensitive to contextual changes (Sclar et al., 2024), so having a conversation may make an 'persona drift' during the interaction. A single case study on GPT (Frisch and Giulianelli, 2024) supports this claim: expressed persona can be changed only with a few interactions. Despite the case study, the result cannot be easily generalized to other models due to the difference in model families and sizes. Hence, we need a study to identify model-specific differences.

Thus, this paper compares the patterns of persona drift across nine LLMs and attempts to reveal the cause of such drifts. Especially, as our motivation begins with the persona of chatbots, we wanted to know whether LLMs suffer persona drifts during a conversation. In the experiment, we asked two LLM agents to discuss 36 themes that are related to one's life, emotions, values, and feelings. We borrowed these themes from human study (Aron et al., 1997) since they make humans discuss their personality. After collecting conversations, we analyze persona drift with the following two questions.

RQ1. Does LLM's architecture affect persona drift?

This research question focuses on the effect of model structure. As parameter sizes and model families may affect the performance and behavior of

LLMs, we also suspect that such differences can cause changes in persona drifts. Thus, we employ a systematic comparison. Using topic modeling and PsychoBench (Huang et al., 2023), we successfully identified a relationship between model structure and persona drift. Here, we used an empty given persona because we wanted to focus on the effect of model structure, rather than that of given persona.

RQ2. Does given persona affect persona drift?

We pose another research question to observe the effect of given persona. Specifically, we provide two types of persona inputs based on how strongly the prompt encourages the LLM to be influenced by the conversational partner: high-sensitive and low-sensitive. As instruction-tuned LLMs generally follow the prompt faithfully, we hypothesize that high-sensitive personas, which are influenceable and empathetic, will generate expressed personas less consistently than low-sensitive personas.

2 Related Work

Researchers have been examining two factors that affect consistency in conversations: memory and expressed persona. Because people generally expect consistency throughout a dialogue, researchers first started by examining memory consistency, which can easily form a task. A large body of existing research has focused on how memory is retained, largely verifying whether an LLM continues to remember certain information during conversation (Tseng et al., 2024; Chen et al., 2023; Maharana et al., 2024; Zhang et al., 2024b; Afzoon et al., 2024). For instance, Chen et al. (2023) analyzed how consistently an LLM can uphold a given memory. Meanwhile, Maharana et al. (2024) created the LoCoMo dataset to investigate how well they remember information over prolonged conversations.

However, memory is not the only factor that affects task performance or the naturalness of a dialogue; persona should be given (Wu et al., 2023; Li et al., 2023; Abbasiantaeb et al., 2024; Zhang et al., 2024a). For example, Abbasiantaeb et al. (2024) reported that it is possible to model a conversational question-answering task as a conversation between a teacher and a student using an LLM. By qualitatively assessing the quality of the interaction, they found that providing two identities could improve the interaction process in a more human-like manner. Similarly, Li et al. (2023) simulated a job fair scenario with two agents: a job seeker and an employer. They explored how their cooperative interaction affects task performance. However, these studies assume that LLMs express the same persona consistently when a conversation progresses. Considering that the memory changes during a conversation, expressed persona could also change.

Hence, recently, researchers attempted to quantify the expressed persona before measuring its consistency. Some researchers designed benchmarks measuring the persona of LLM (Huang et al., 2023; Wang et al., 2024; Zhang et al., 2024a; Frisch and Giulianelli, 2024). For example, Huang et al. (2023) assessed the persona of LLMs using fourteen types of questionnaires. Though they found that different LLMs exhibit different identities, they did not measure the impact of conversation on the persona. However, measuring the impact is crucial because accumulated chat histories can introduce unexpected changes, as memory-related studies suggested. Frisch and Giulianelli (2024) supports this claim with a case study. They demonstrated that GPT tends to adopt each other's persona in an interaction, failing to maintain persona. Though this paper addressed the problem we call persona drift, it has some limitations when applied to conversational agents; the interaction was unidirectional, which is different from a usual conversation, as they asked agents to continue writing others' work. So, it is yet unanswered whether LLMs can consistently maintain its persona expression in a bidirectional conversation. We suspect a bidirectional conversation may cause different tendencies in persona drift compared to a unidirectional one.

3 Experiments

To investigate factors influencing persona drift issue of LLMs, we conduct an experiment[1]. We ask two LLM agents to discuss 36 themes. During the conversation, we collect their conversation logs and measure expressed persona based on the conversation. Using both qualitative and quantitative analyses, we attempt to answer two research questions on persona drift. Thus, this section first describes LLM agents used (Sections 3.1 and 3.2). Next, we explain experimental procedure (Section 3.3) and analysis methods (Sections 3.4 and 3.5).

3.1 RQ1: Language Models Tested

We compared nine models, considering their popularity, parameter size, and architecture. Based on popularity, we selected GPT, the most famous black

[1]Code is available at [blinded for review].

Family	Parameter Sizes		
	Small	Medium	Large
LLaMA 3.1	8B	70B	405B
Mixtral	8x7B	8x22B	
Qwen 2	7B	72B	
GPT	*Undisclosed*: 3.5 Turbo, 4o		

Table 1: Models tested in our experiment

box LLM, and three famous open-sourced families: LLaMA, Mixtral, and Qwen. Table 1 shows the nine models with their parameter sizes[2]. According to parameter sizes, we partitioned open-sourced models into three categories: small (models with $<$ 20 billion parameters), medium ($<$ 100B), and large ($\geq$ 100B). This categorization allows a systematic comparison of model characteristics based on parameter scale. We did not assign GPT models into any size groups since their parameter sizes were not officially disclosed. To focus on the effect of model itself, here we did not provide any persona-related information in the input prompt.

GPT This family comprises GPT-3.5 Turbo (Brown et al., 2020) and GPT-4o (Hurst et al., 2024). Although their parameter sizes remain undisclosed, these models were included in the experiment due to their high performance and widespread recognition in practice.

LLaMA3.1 This family includes LLaMA 3.1-8B, 3.1-70B, and 3.1-405B (Dubey et al., 2024). While sharing the same basic architecture, they differ substantially in parameter size.

Mixtral This family contains Mixtral8x7B and 8x22B (Jiang et al., 2024). It employs a Mixture-of-Experts (MoE) architecture, which differs from other open-sourced models. Thus, comparing Mixtral with others can hint at how MoE influences expressed persona.

Qwen This family encompasses Qwen2 7B and Qwen2 72B (Yang et al., 2024). Advertised as particularly adept at conversational tasks, these models may provide a hint for how expressed persona changes in a conversation.

3.2 RQ2: Given personas

After investigating RQ1, we examine the effect of given persona. As we suspect the effect is not large enough to offset the effect of model-related factors, we used two LLMs whose persona drifts are the most severe among the nine models. Though users expect LLMs can maintain consistent persona expression, those two models should maintain its persona to meet the expectation.

As LLMs are trained to follow instructions, we suspect that LLMs can be affected by how well a conversation influences the given persona. That is, LLMs may exhibit greater persona drift when they operate under an input prompt asking for higher sensitivity. Accordingly, we define two persona groups: (1) the high-sensitive group and (2) the low-sensitive group. High-sensitive personas are characterized by elevated emotional sensitivity and empathy. Humans in this group adjust their responses and self-presentation more flexibly during the conversation (Davis, 1983; Brennan, 1998; Dietz and Kleinlogel, 2014), and we believe that LLMs learned this behavior. In contrast, low-sensitive personas are defined as those who are outgoing and goal-oriented, with lower emotional sensitivity. Humans in this group reveal a more stable responses despite external conversational influences (John et al., 1999; Su et al., 2019; Locke and Latham, 2002; Gross and John, 2003).

To systematically investigate persona drift, we designed 20 distinct persona profiles for each group, balanced for gender and age distribution (18-29 years). Each persona was structured around four key components (Personality, Interpersonal Relationship, Motivation, and Emotion) with specific trait configurations to ensure internal consistency within each influence type. For example, high-sensitive personas combined emotionally sensitive personality traits with attachment-oriented relationship styles, while low-sensitive personas paired goal-oriented traits with assertive communication patterns. This design allowed us to analyze how different persona configurations affect its consistency during extended conversations. Appendix B.5 explains the detailed setup.

3.3 Procedure for Generating conversation

Our generation procedure adopts a procedure of a psychological study (Aron et al., 1997) that makes participants have a deep conversation about themselves. Despite the study examining different dependent variables from ours, we chose the study for two reasons. First, the method suggests a scientific way to identify changes during a conversation. They let humans have a conversation about 36

[2]We assigned Mixtral by active parameters (13B and 39B), from `https://mistral.ai/en/news/mixtral-8x22b`.

themes and measured human psychological states three times within the conversation. By comparing three measured values, they could statistically identify the changes. As we also aimed to measure changes in persona, we borrowed their setup.

Second, the method uses materials that are highly related to identity of someone. The 36 themes used in the study directly or indirectly ask participants to answer their thoughts about their lives, values, or motivations. So, it is highly likely that the answer contains concepts related to their identity. In the view of LLMs, such answers may ignite some related tokens during the generation procedure. That is, expressed persona may be easily affected by the words in the previous discussion, which are due to the contextual adaptation of LLMs. Nonetheless, LLMs should not exhibit persona drift because humans expect personas to seldom change in a short context. Thus, we adopted the study.

In the generation procedure, we asked two agents answer the 36 themes in Aron et al. (1997). For each theme, we pose a question about the theme. Then, one agent generates a response, considering previous chat history. Then, the other agent generates response to the question, considering previous history and the first agent's response. We repeated this procedure until the end of 36 themes and collected conversation logs to answer two RQs. For RQ1, we simulated 20 conversations for each LLM. For RQ2, we simulated 10 conversations for each persona group: we paired similar personas to avoid its drift effect reported by Frisch and Giulianelli (2024). To obtain diverse conversation logs, we set the temperature parameter at 0.7[3]. As a result, we gathered 400 logs for each RQ.

3.4 Qualitative: Topic modeling

As a qualitative analysis, we employed a topic modeling method, BERTopic (Grootendorst, 2022). The unit of analysis for the topic exploration was a single utterance, defined as one agent's response to one of the 36 themes. So, we used 12,960 utterances[4] for the topic analysis in total. We post-processed these utterances by applying stop-word removal, and extracted topics which contains more than 50 support utterances.

To discover differences across architectures including model size or family, we grouped models into several groups and conducted topic analyses for each group. Comparing differences in topics may provide insights into how different model architectures influence expressed persona. For parameter sizes, we ran topic modeling for each parameter size group: small, middle, and large. Also, we ran topic modeling for each model family: GPT, LLaMA, Mixtral, and Qwen. After topic modeling, we chose the ten most representative topics from each topic model. We manually associated each topic with one of the 36 predefined themes by tracing back to the corresponding conversation logs and identifying the most frequent theme that the topic frequently appeared in.

Similarly, we also extracted topics for each persona group. We separately extracted topics for high-sensitive and low-sensitive identities for RQ2. Procedure for analyzing topics are the same as the method used for analyzing model architectures.

3.5 Quantitative: PsychoBench and MFQ

As a quantitative analysis, we used PsychoBench (Huang et al., 2023) and Mcgill's Friendship Questionnaire (MFQ; Mendelson and Aboud (1999)). These artifacts can measure expressed persona. PsychoBench contains thirteen questionnaires from psychology, quantifying four parts of one's persona: personality, interpersonal relationship, motivation, and emotion. We expect these four parts keep unchanged during a conversation. MFQ quantifies how one thinks about the conversational partner. We included this questionnaire to track how the conversational agents think each other. Detailed descriptions for questionnaires are in Appendix A.

We measured those questionnaires three times within a conversation. Inspired by Aron et al. (1997), we set three snapshots for each conversation: after answering 12th, 24th, and 36th themes. Then, we applied PsychoBench and MFQ on those snapshots. As in PsychoBench, we asked LLMs to answer the questionnaire ten times at temperature zero, with randomized order of questions for each run to mitigate primacy effects (Wang et al., 2023). The only difference between our method and PsychoBench is that we measured persona with previous conversation history, rather than measuring without the history. This approach allows us to capture dynamic persona changes during interaction. As a result, we could obtain three intermediate scored responses for each conversation.

Using the scored responses, we performed statis-

[3] We set the temperature to 0.7 for all experiments to ensure consistency, since 0.7 was the default temperature value when we ran the experiment.

[4] 12,960 = 20 conversations × 2 agents × 36 themes × 9 models

Small-sized open-source models (≤ 10B)	Theme
#0 friendship, trust, respect, mutual, means	20
#1 *users*, language, accomplishments, accomplishment, *assist*	(AI)
#2 feel, way, appreciate, grateful, admire	31
#3 regret, told, expressing, having, feelings	33
#4 dont, *digital*, exist, existence, designed	(AI)
#5 shared, understanding, conversations, mutual, deep	20
#6 death, living, live, die, hunch	7
#7 rehearsing, rehearse, ensure, helps, especially	3
#8 humor, topics, jokes, issues, sensitive	32
#9 singing, sang, sing, karaoke, fun	5
Middle-sized open-source models (10B - 100B)	**Theme**
#0 way, really, appreciate, feel, qualities	31
#1 know, friendship, honesty, value, want	20
#2 statements, shared, value, growth, conversations	25
#3 regret, told, having, loved, ive	33
#4 languages, ability, cultures, language, speak	12
#5 living, die, focusing, present, healthy	7
#6 childhood, family, happy, warm, close	23
#7 fascinating, conversation, choose, elon, musk	1
#8 accomplishment, greatest, hard, proud, achievement	15
#9 mother, relationship, shes, guidance, loving	24
Large-sized open-source models (> 100B)	**Theme**
#0 statements, friendship, life, having, grateful	20
#1 ive, accomplishment, life, greatest, encouraged	11
#2 really, way, youre, feel, like	31
#3 regret, told, having, ive, think	33
#4 live, left, focus, try, make	19
#5 feeling, ive, youre, problem, advice	36
#6 embarrassing, memory, ended, moment, painful	29
#7 affection, love, relationship, mother, believe	21
#8 id, able, famous, ability, language	12
#9 know, want, im, id, bit	27

Table 2: Top 10 topics discovered per parameter size groups. Italicized words indicate 'as an AI' response. Underlined words are related to pronouns.

tical tests. First, we verify whether expressed persona changed during a conversation. We used the repeated measure ANOVA or Friedman test (Girden, 1992; Friedman, 1937), regarding normality of scored responses. Second, we checked consistency by conducting post-hoc tests: Tukey's test or Wilcoxon signed-ranked test (Tukey, 1949; Woolson, 2005), regarding normality. To mitigate potential type I errors arising from multiple comparisons, we used Bonferroni correction to adjust p-values conservatively in Wilcoxon test (Bonferroni, 1936).

4 Result and Discussion

4.1 RQ1: Effect of Structure

The result for RQ1 indicates that the effect of model-related factor exists. Specifically, parameter sizes showed a large impact on consistency. The effect of model family is lower than that of the size.

Effect of parameter sizes: According to the qualitative analysis, two notable changes were observed in the representative topics among different parameter sizes: those pertaining to "AI" and to "pronouns." The result is shown in Table 2. First, regarding AI, small LLMs refuse to engage in conversations on a given theme as they are an AI. As shown in Topics #1 and #4 for the small models, they tended to refuse or guard their own responses. This tendency was not observed in the medium or large models. That is, though the safeguard was strongly activated in small models, those of middle or large models were less strong.

Second, regarding pronouns, large LLMs generates its responses based on fictitious information about itself or the other participant. Here, we define fictitious information as a falsy plausible memory of LLMs about their non-existing human life or historical memories, which is slightly different from hallucinations about the existing facts. Though most pronouns were filtered as stop-words in topic modeling, some pronoun-based words are discovered; for example, "I've" in Topic #3 of Middle group. We observed that these pronouns were usually used to indicate fictitious person in LLMs' falsy memory to create a plausible story. Compared to the small models (0 pronouns), medium and large models (2 and 8 pronouns) used pronouns more frequently. Due to the recency effect and other biases in LLMs, such fictitious contents may influence subsequent conversations. This claim is also supported by themes co-occurring across size groups. For example, Theme 31, which asks about one's perception of the other participant, appears in all size groups. But, only the large models used second-person pronouns referring to the other participant (Large #2). Similar phenomenon also happens in Theme 33, which asks about one's regrets.

The quantitative result supports these observations; as the parameter size increases, LLMs exhibit more persona drifts, as shown in Table 3. The small models show the best consistency of persona, while the total count of consistent factors decreases on larger models. LLaMA model clearly shows this tendency, where the total count sharply decreases. Mixtral and Qwen families show similar patterns.

Combining these results indicates that larger models tend to introduce fictitious information, making it suffer persona drifts. Large models introduce fictitious details about themselves. So, those LLMs receive new fabricated information as credible source of their persona. Consequently, such

	Conditions:	*Without* any given persona									*With* a given persona			
	Family:	GPT		LLaMA 3.1			Mixtral		Qwen 2		GPT-4o		L 405B	
		3.5T	4o	8B	70B	405B	7B	22B	7B	72B	low	high	low	high
(1) Personality														
BFI	*Openness*				✓		✓	✓	✓				✓	
	Conscientiousness				✓		✓	✓	✓	✓			✓	
	Extraversion					✓	✓		✓	✓			✓	
	Agreeableness			✓	✓		✓	✓	✓	✓				
	Neuroticism				✓		✓		✓	✓			✓	✓
EPQ-R	*Extraversion*							✓	✓	✓				
	Psychoticism							✓	✓	✓				
	Neuroticism							✓	✓	✓				
	Lying			✓				✓	✓	✓				
DTDD	*Machiavellianism*									✓		✓	✓	
	Psychopathy			✓			✓		✓	✓		✓	✓	
	Narcissism			✓			✓		✓	✓		✓		
	Consistent factors (12)	0	0	4	4	1	7	7	11	11	0	3	6	1
(2) Interpersonal Relationship														
BSRI	*Masculine*		✓	✓			✓		✓	✓				✓
	Feminine						✓	✓	✓					✓
CABIN	*Realistic*	✓		✓			✓	✓					✓	
	Investigate	✓		✓		✓	✓	✓					✓	
	Artistic		✓	✓			✓	✓					✓	
	Social	✓		✓			✓	✓					✓	
	Enterprising	✓	✓	✓			✓	✓					✓	
	Conventional	✓		✓			✓	✓					✓	
ICB	*Overall*		✓			✓	✓	✓	✓		✓		✓	✓
ECR-R	*Attachment Anxiety*	✓		✓						✓		✓		
	Attachment Avoidance			✓			✓	✓		✓		✓		
MFQ	*Stimulating companionship*			✓			✓		✓					
	Help			✓			✓		✓					
	Intimacy			✓			✓		✓					
	Reliable alliance			✓			✓		✓					
	Self-validation			✓			✓		✓					
	Emotional security			✓			✓							
	Consistent factors (17)	6	4	15	0	2	16	9	8	3	1	2	7	3
(3) Motivation														
GSE	*Overall*	✓	✓			✓		✓						
LOT-R	*Overall*					✓	✓	✓			✓			
LMS	*Rich*			✓						✓				
	Motivator												✓	
	Important									✓			✓	
	Consistent factors (5)	1	1	1	0	2	1	2	0	2	1	0	2	0
(4) Emotion														
EIS	*Overall*			✓			✓							✓
WLEIS	*Self-emotion appraisal*			✓	✓		✓							✓
	Others' emotion appraisal				✓									✓
	Use of emotion													✓
	Regulation of emotion					✓	✓							✓
Empathy	*Overall*			✓		✓		✓	✓			✓	✓	✓
	Consistent factors (6)	0	0	3	2	2	3	1	1	0	0	1	1	6

Table 3: Verification of whether expressed persona was retained during the conversation for each subscale. Checkmarks (✓) indicate the persona change is statistically insignificant in both Friedman and posthoc tests. Detailed statistical results are shown in Appendix (Tables from 10 to 13).

GPT family	Theme	LLaMA 3.1 family	Theme
#0 thoughtful, admire, genuine, appreciate, empathy	28	#0 dont, personal, information, *assist*, provide	(AI)
#1 enjoy, value, meaningful, growth, appreciate	8	#1 desire, value, nature, conversations, based	25
#2 value, friendship, honesty, important, trust	27	#2 way, really, feel, youre, like	31
#3 regret, told, expressing, feelings, telling	33	#3 regret, told, having, ive, ones	33
#4 youd, discuss, free, like, im	(AI)	#4 famous, id, author, music, renowned	2
#5 affection, love, emotional, play, belonging	21	#5 friendship, means, having, accepts, connection	20
#6 greatest, accomplishment, far, completing, overcoming	15	#6 rehearse, helps, avoid, ensure, yes	3
#7 ability, choose, wake, tomorrow, speak	12	#7 da, leonardo, vinci, facinating, art	1
#8 year, knew, focus, left, prioritize	19	#8 singing, sang, favorite, driving, ago	5
#9 means, friendship, having, trust, mutual	20	#9 topics, joked, humor, issues, hurtful	32

Mixtral family	Theme	Qwen family	Theme
#0 appreciate, admire, humor, feel, kindeness	31	#0 *ai*, dont, *users*, *assist*, information	(AI)
#1 live, living, make, time, die	19	#1 kindness, qualities, admire, humor, thoughtful	31
#2 told, regret, expressing, having, express	33	#2 living, focusing, time, experiences, death	7
#3 accomplishment, greatest, life, career, work	11, 15	#3 impact, world, accomplishment, positive, career	13
#4 statements, shared, value, importance, enjoy	25	#4 shared, interests, committed, statements, learning	25
#5 *users*, language, *model*, *artificial*, *ai*	(AI)	#5 regret, expressing, gratitude, feelings, loved	33
#6 humor, topics, mindful, jokes, joking	32	#6 honesty, respect, friendship, mutual, value	16
#7 dinner, obama, michelle, guest, choice	1	#7 loss, disturbing, losing, profoundly, profound	35
#8 day, perfect, relaxation, involve, activities	4	#8 languages, cultures, exposure, ability, different	12
#9 mind, body, mental, 30yearold, retain	6	#9 memories, treasured, cherished, sharing, memory	17

Table 4: Top 10 topics discovered per family. Italicized and underlined words indicate 'as an AI' and pronouns.

fictitious details lead to fluctuations in persona. Indeed, after reading the logs, we found a tendency of larger models to make a fictitious details about themselves or conversation partners. For example, they easily describe imaginary aspects of one's own inner world. Small models, in contrast, do not rely on either themselves or the partner; rather, we found that they strive to thoroughly explain given concepts. See Appendix C for representative examples. So, these smaller models do not generate emotional matters that could influence persona expressions, leading to a relatively stable persona in Table 3. However, we should also keep in mind that small models just explain the concept as an AI, rather than engaging in the chat.

Effect of model families: According to the qualitative analysis, slight differences in topics were observed among the models. Table 4 shows the result. Similar to parameter sizes, we focused on two aspects: AI and pronouns. First, regarding AI, all models exhibit a topic to refuse answers as an AI: GPT #4, LLaMA #0, Mixtral #5, and Qwen #0. Second, pronouns appear only in GPT (2 pronouns) and LLaMA (3 pronouns), but not in Mixtral or Qwen though the difference is not large.

The quantitative analysis yields similar findings, suggesting that only slight differences exist among the models. Comparing each model series in Table 3 reveals that Mixtral and Qwen maintain persona well in certain parts of identity. In particular, Qwen can maintain personality in most cases, while Mixtral consistently retains interpersonal relationship aspects. In contrast, GPT and LLaMA families generally struggle to maintain persona.

In summary, parameter size has a stronger influence on persona drift than model families. Although we could observe certain distinctions within the Mixtral and Qwen families, their impact seems limited to specific models. In contrast, parameter size consistently affects all four models, often causing larger drifts. Thus, we concluded that parameter size is a more significant factor to build a consistent persona expression than model families.

4.2 RQ2: Effect of given persona

The experimental results for RQ2 indicate that the model-related effect is stronger than the effect of given persona. In this section, we describe the result along two main dimensions: (1) comparison between LLMs without persona (RQ1) and LLMs with given personas (RQ2), and (2) comparison between high- and low-sensitive persona. Note that we used GPT-4o and LLaMA 3.1 405B for RQ2, as they are two models whose persona drift is large.

In the following subsections, we focus primarily on describing overall tendencies rather than definitive possible causal factors. Because of two obstacles, we could not identify possible causes. First, though we conducted a topic analysis, we found no significant differences among the groups. So, we decided to illustrate topics in the Appendix C instead of analyzing here. Second, due to the black-box nature of GPT-4o, it is hard to identify any explanations about the difference.

4.2.1 Impact of Given Persona

Our experiment shows that the influence of the model family appears to be greater than that of the given persona. The last four columns in Table 3 show the result. Comparing the results of the persona-assigned models with results from RQ1, we observe that GPT-4o still struggles to maintain a consistent persona expression. In the case of GPT-4o without a given persona, expressed persona was retained across five factors in total. However, even when a persona was given, only two factors in the low-sensitive category and six factors in the high-sensitive category were consistently maintained, indicating that the model's ability to preserve persona expression does not significantly improve with explicit persona assignment. In contrast, LLaMA3.1 405B demonstrates the ability to retain persona expression in certain factors. In RQ1, LLaMA3.1 405B maintained expressed persona across seven factors in total. However, when we assign a persona, the model retained persona expression in 16 factors in the high-sensitive category and 10 factors in the low-sensitive category. This suggests that LLaMA can maintain persona expression in specific factors, though it can not maintain consistency of the whole identity. Hence, we conclude that assigning a persona does not necessarily guarantee consistency of expressed persona; the result may vary across models.

4.2.2 Impact of Persona Sensitivity

As we concluded that the model difference has a greater impact than the given persona, here we discuss the effect of given persona for each LLM separately. First, the GPT-4o model generally struggles to maintain the expressed persona, regardless of the type of persona given. Table 3 shows that GPT-4o achieves more consistency in high-sensitive (0, 1, 1, and 0 factors for each part) compared to low-sensitive (3, 2, 0, and 1 factors). Specifically, GPT-4o retained factors related to emotional influence, including attachment or empathy. The model also retained persona expression on DTDD questionnaire, which are related to dark personality factors: one's willingness to control others. We suspect this phenomenon is because given personas instruct GPT-4o to follow other's emotions.

Second, LLaMA 3.1 405B exhibits a different pattern; LLaMA preserves persona expressions more in low-sensitive personas. Specifically, the model with a low-sensitive persona tends to retain identity in two parts: personality (6 factors) and interpersonal relationships (7 factors). Meanwhile, the model with a high-sensitive persona shows a stronger tendency to maintain the emotional part of the identity (6 factors), which is similar to the case of GPT-4o. Hence, we suspect that certain parts of the identity are more likely to be preserved depending on the interaction effect between model family used and given persona type fed to the model.

5 Conclusion

This study examined whether LLMs can maintain its persona expression in long-term conversations. We also wanted to identify the effect of parameter sizes, model families, and given persona on maintaining its persona expression. So, we set two research questions. First, we investigated whether LLMs could maintain consistent interaction patterns (which we call expressed persona) without a given persona in the input prompt. We qualitatively analyzed logs of 36-turn conversations and statistically verified the research question. Second, we conducted the same experiment while we gave a specific persona as an input into LLMs. We analyzed the difference between LLMs without any given persona, those with low-sensitive persona, and those with high-sensitive persona.

As a result, we found three things. First, regarding the parameter sizes, larger models exhibited greater persona drift and struggled more with maintaining a stable persona expression than smaller models. Second, regarding the model families, the effect of the model family is relatively smaller than the effect of the parameter sizes, though we observed some differences across models. Third, regarding persona assignment, the assignment alone does not ensure consistency of expressed persona; rather, the model's inherent characteristics play a greater role in determining how well it maintains its persona expressions throughout the conversation. Overall, these results highlight the challenges of maintaining consistent persona expression in LLM-based dialogues, emphasizing the need for further research on model-specific analysis or strategies for maintaining persona. We believe this study can lay a cornerstone for understanding how LLMs handle a given persona and its expression.

Limitation

This work has four limitations when applying our findings to other studies. First, while we aimed to encourage open-ended responses, conversations followed structured themes to obtain coherence across multiple runs. As a result, questions were introduced to guide the dialogue, limiting full free-form interaction. Although this approach was necessary for maintaining a meaningful conversational flow, it may have influenced the natural development of persona drift or expressions of the persona.

Second, though our analysis focused on whether an LLM maintains its given persona, we did not examine the detailed dynamics of how individual identity factors fluctuate over time. Understanding the specific aspects of persona drift, such as variations in emotional consistency or interpersonal parts, requires further investigation to deepen our comprehension of persona drift in LLMs.

Third, although we identified persona drift, we did not propose specific methods for controlling or mitigating it through prompt engineering or model adjustments. Future research should explore intervention strategies to stabilize persona expressions and assess effectiveness in long-term interactions.

Fourth, we tested LLMs with a simple set of persona descriptions. If given persona descriptions contain more detailed or descriptive information, different outcomes might emerge. The impact of persona complexity on persona drift remains an open question, warranting further exploration to assess how variations in persona richness influence conversational consistency.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Saleh Afzoon, Usman Naseem, Amin Beheshti, and Zahra Jamali. 2024. Persobench: Benchmarking personalized response generation in large language models. *arXiv preprint arXiv:2410.03198*.

Arthur Aron, Edward Melinat, Elaine N Aron, Robert Darrin Vallone, and Renee J Bator. 1997. The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and social psychology bulletin*, 23(4):363–377.

Sandra L Bem. 1974. The measurement of psychological androgyny. *Journal of consulting and clinical psychology*, 42(2):155.

Sandra Lipsitz Bem. 1977. On the utility of alternative procedures for assessing psychological androgyny. *Journal of consulting and clinical psychology*, 45(2):196.

C.E. Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber.

KA Brennan. 1998. Self-report measurement of adult attachment: An integrative overview. *Attachment theory and close relationships/Guilford*.

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.

Melody Manchi Chao, Riki Takeuchi, and Jiing-Lih Farh. 2017. Enhancing cultural intelligence: The roles of implicit culture beliefs and adjustment. *Personnel Psychology*, 70(1):257–292.

Ruijun Chen, Jin Wang, Liang-Chih Yu, and Xuejie Zhang. 2023. Learning to memorize entailment and discourse relations for persona-consistent dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 12653–12661.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Joerg Dietz and Emmanuelle P Kleinlogel. 2014. Wage cuts and managers' empathy: How a positive emotion can contribute to positive organizational ethics in difficult times. *Journal of business ethics*, 119:461–472.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Sybil Bianca Giuletta Eysenck, Hans Jürgen Eysenck, and Paul Barrett. 1985. A revised version of the psychoticism scale. *Personality and Individual Differences*, 6:21–29.

R Chris Fraley, Niels G Waller, and Kelly A Brennan. 2000. An item response theory analysis of self-report measures of adult attachment. *Journal of personality and social psychology*, 78(2):350.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 102–111.

ER Girden. 1992. *ANOVA: Repeated measures*, volume 84. Sage.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

James J Gross and Oliver P John. 2003. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of personality and social psychology*, 85(2):348.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Peter K Jonason and Gregory D Webster. 2010. The dirty dozen: a concise measure of the dark triad. *Psychological assessment*, 22(2):420.

Yuan Li, Yixuan Zhang, and Lichao Sun. 2023. Metaagents: Simulating interactions of human behaviors for llm-based task-oriented coordination via collaborative generative agents. *arXiv preprint arXiv:2310.06500*.

Edwin A Locke and Gary P Latham. 2002. Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American psychologist*, 57(9):705.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Morton J Mendelson and Frances E Aboud. 1999. Measuring friendship quality in late adolescents and young adults: Mcgill friendship questionnaires. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, 31(2):130.

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*.

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*.

Michael F Scheier and Charles S Carver. 1985. Optimism, coping, and health: assessment and implications of generalized outcome expectancies. *Health psychology*, 4(3):219.

Michael F Scheier, Charles S Carver, and Michael W Bridges. 1994. Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): a reevaluation of the life orientation test. *Journal of personality and social psychology*, 67(6):1063.

Nicola S Schutte, John M Malouff, Lena E Hall, Donald J Haggerty, Joan T Cooper, Charles J Golden, and Liane Dornheim. 1998. Development and validation of a measure of emotional intelligence. *Personality and individual differences*, 25(2):167–177.

R Schwarzer. 1995. Generalized self-efficacy scale. *Measures in health psychology: A user's portfolio. Causal and control beliefs/Nfer-Nelson*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Rong Su, Louis Tay, Hsin-Ya Liao, Qi Zhang, and James Rounds. 2019. Toward a dimensional model of vocational interests. *Journal of Applied Psychology*, 104(5):690.

Thomas Li-Ping Tang, Toto Sutarso, Adebowale Akande, Michael W Allen, Abdulgawi Salim Alzubaidi, Mahfooz A Ansari, Fernando Arias-Galicia, Mark G Borg, Luigina Canova, Brigitte Charles-Pauvers, et al. 2006. The love of money and pay level satisfaction: Measurement and functional equivalence in 29 geopolitical entities around the world. *Management and Organization Review*, 2(3):423–452.

Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate chatbots to facilitate critical thinking on youtube: Social identity and conversational style make a difference. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–24.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Raphael Vallat. 2018. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023. Primacy effect of chatgpt. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

Chi-Sum Wong and Kenneth S Law. 2017. The effects of leader and follower emotional intelligence on performance and attitude: An exploratory study. In *Leadership perspectives*, pages 97–128. Routledge.

Robert F Woolson. 2005. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics*, 8.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024a. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024b. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

A Explanation for Used Questionnaires

As the experiment requires measuring 15 questionnaires on each snapshot of conversation, we modified the PsychoBench framework by Huang et al. (2023) to measure psychological states on each snapshot. So, we employed 14 questionnaires in PsychoBench and added MFQ to measure how LLM perceives the conversational partner as a factor in the interpersonal relationship aspect. To help readers understand, we further elaborated on those 15 psychological questionnaires regarding their goals and included factors.

A.1 Personality

Big Five Inventory (BFI) is a widely-used questionnaire to measure one's personality across five key dimensions(John et al., 1999). First, an increase in *openness* suggests the agent becomes more inventive and curious about a new experience. Second, an increase in *conscientiousness* suggests the agent becomes more efficient and organized when doing a task. Third, an increase in *extraversion* suggests the agent shows more outgoing and energetic behaviors. Fourth, an increase in *agreeableness* suggests the agent becomes more friendly and compassionate to the others. Lastly, an increase in *neuroticism* suggests the agent becomes more emotionally sensitive and nervous to a stressor.

Eysenck Personality Questionnaire, Revised (EPQ-R) is a questionnaire that attempts to identify individual differences in temperament and behavior(Eysenck et al., 1985). This questionnaire is commonly used in clinical and psychological research, and it has four factors. First, an increase in *extraversion* suggests the agent becomes more outgoing, talkative, and needs external stimulation. Second, an increase in *neuroticism* suggests the increment in the levels of negative affections, including depression and anxiety. Third, an increase in *psychoticism* suggests the agent expresses more aggressive behaviors and is more likely to show a psychotic episode or symptoms. Lastly, an increase in *lying* suggests the agent becomes more likely to make a lie or dissimulate to satisfy its social desirability.

Dark Triad Dirty Dozen (DTDD) is a clinical questionnaire measuring the possible presence of three dark traits(Jonason and Webster, 2010). First, an increase in *machiavellianism* suggests the agent becomes more likely to manipulate others, show indifference to morality, and focus on its own interest. Second, an increase *narcissism* suggests the agent shows a more excessive preoccupation with itself and its own needs, even when it needs to sacrifice others. Lastly, an increase in *psychopathy* suggests the agent shows more egocentric and bold behaviors combined with impaired empathy.

A.2 Interpersonal Relationship

Bem's Sex Role Inventory (BSRI) is a questionnaire about how the agent identifies itself psychologically regarding two gender roles(Bem, 1974, 1977). An increase in *masculinity* suggests the agent becomes more assertive, ambitious, competitive, and dominant. Meanwhile, an increase in *femininity* suggests the agent becomes more affectionate, cheerful, and childlike.

Comprehensive Assessment of Basic Interests (CABIN) is a questionnaire about an individual's basic interest(Su et al., 2019). This measures one's preferences in 41 domains from six categories. We used the six categories in our experiment. First, agents with high *realistic* category favor practical or hands-on experiences. Second, agents with high *investigative* category prefer scholastic or intellectual opportunities. Third, agents with high *artistic* category favor creative and expressive experiences. Fourth, agents with high *social* category prefer to work with others to help them grow. Fifth, agents with high *enterprising* category favor opportunities in leading or managing people. Lastly, agents with high *conventional* category prefer routine and well-structured environments.

Implicit Culture Belief (ICB) is a questionnaire about the effect of implicit ethnic cultural influences on one's belief(Chao et al., 2017). High *overall* score in this questionnaire indicates high cultural influences in the agent's belief.

Experiences in Close Relationships, Revised (ECR-R) is a questionnaire about an adult's attachment in a romantic relationship(Fraley et al., 2000; Brennan, 1998). This measures two forms of insecure attachments. First, agents with high *attachment anxiety* worry that they will become estranged from their partners. Second, agents with high *attachment avoidance* try to keep psychological distance from their partners.

McGill Friendship Questionnaire - Friend's Function (MFQ-FF) is a questionnaire about how the agent perceives the function of its partner(Mendelson and Aboud, 1999). This questionnaire is different from other interpersonal relationship questionnaires because it assumes the presence of a specific partner; the response is based on the agent's thoughts about that partner. MFQ has six factors. First, an agent answering high *stimulating companionship* perceives he can do enjoyable or exciting things with his partner. Second, an agent answering high *help* thinks that his partner is good at providing guidance or assistance. Third, an agent answering high *intimacy* thinks that his partner is sensitive to his needs and states and open to honest expressions of thoughts. Fourth, an agent answering high *reliable alliance* regards his partner as an always available and loyal friend. Fifth, an agent answering high *self-validation* thinks his partner encourages and helps him maintain a positive self-image. Lastly, an agent answering high *emotional security* thinks his partner provides comfort and confidence in a novel situation.

A.3 motivation

General Self-Efficacy (GSE) is a questionnaire about one's perceived efficacy for coping with a situation, performing a task, and achieving goals(Schwarzer, 1995). Agents with high *overall* scores have a high level of self-efficacy; that is, they perceive themselves as good at coping with a difficult situation and achieving goals.

Life Orientation Test, Revised (LOT-R) is a questionnaire about how optimistic or pessimistic the agent perceives about the future (Scheier et al., 1994; Scheier and Carver, 1985). Agents with high *overall* scores expect their future in an optimistic way.

Love of Money Scale (LMS) is a questionnaire about one's attitude toward money and financial incentives through three factors (Tang et al., 2006). First, an increase in *rich* suggests the agent has more positive feelings towards money. Second, an increase in *motivator* suggests the agent becomes more easily motivated by monetary incentives. Third, an increase in *important* suggests the agent has a stronger belief that money means power, freedom, security, or other important values.

A.4 Emotion

Emotional Intelligence Scale (EIS) is a questionnaire measuring one's emotional intelligence (Schutte et al., 1998). Agents with high *overall* scores have a strong understanding and control of their emotions.

Wong and Law Emotional Intelligence Scale (WLEIS) is a questionnaire about emotional intelligence in the workplace, regarding four factors (Wong and Law, 2017). First, agents with high *self-emotion appraisal* can appraise their own emotions. Second, agents with high *others' emotion appraisal* can appraise and recognize the emotions of others. Third, agents with high *use of emotion* use emotions to facilitate performance. Lastly, agents with high *regulation of emotion* can regulate emotions to promote emotional and intellectual growth.

Empathy Scale (Empathy) is a questionnaire about the ability to understand and share the feelings of others. Agents with high *overall* scores can connect with others on an emotional level and respond appropriately to their needs.

B Experimental detail

B.1 36 Conversational Themes

We used 36 conversational themes in the experiment, following Aron et al. (1997). The first 12 themes are used before the first questionnaire measurement.

Theme 1. Given the choice of anyone in the world, whom would you want as a dinner guest?

Theme 2. Would you like to be famous? In what way?

Theme 3. Before making a telephone call, do you ever rehearse what you are going to say? Why?

Theme 4. What would constitute a "perfect" day for you?

Theme 5. When did you last sing to yourself? To someone else?

Theme 6. If you were able to live to the age of 90 and retain either the mind or body of a 30-year-old for the last 60 years of your life, which would you want?

Theme 7. Do you have a secret hunch about how you will die?

Theme 8. Name three things you and your partner appear to have in common.

Theme 9. For what in your life do you feel most grateful?

Theme 10. If you could change anything about the way you were raised, what would it be?

Theme 11. Take 4 minutes and tell your partner your life story in as much detail as possible.

Theme 12. If you could wake up tomorrow having gained any one quality or ability, what would it be?

The next list shows the second 12 themes (from Theme 13 to 24), which are used between the first and the second measurements of questionnaires.

Theme 13. If a crystal ball could tell you the truth about yourself, your life, the future, or anything else, what would you want to know?

Theme 14. Is there something that you've dreamed of doing for a long time? Why haven't you done it?

Theme 15. What is the greatest accomplishment of your life?

Theme 16. What do you value most in a friendship?

Theme 17. What is your most treasured memory?

Theme 18. What is your most terrible memory?

Theme 19. If you knew that in one year you would die suddenly, would you change anything about the way you are now living? Why?

Theme 20. What does friendship mean to you?

Theme 21. What roles do love and affection play in your life?

Theme 22. Alternate sharing something you consider a positive characteristic of your partner. Share a total of 5 items

Theme 23. How close and warm is your family? Do you feel your childhood was happier than most other people's?

Theme 24. How do you feel about your relationship with your mother?

The following is the last list that shows the third 12 themes (from Theme 25 to 36), which are used between the second and the third measurements of questionnaires.

Theme 25. Make 3 true "we" statements each. For instance "We are both in this room feeling..."

Theme 26. Complete this sentence: I wish I had someone with whom I could share...

Theme 27. If you were going to become a close friend with your partner, please share what would be important for him or her to know.

Theme 28. Tell your partner what you like about them; be very honest this time saying things that you might not say to someone you've just met

Theme 29. Share with your partner an embarrassing moment in your life.

Theme 30. When did you last cry in front of another person? By yourself?

Theme 31. Tell your partner something that you like about them already.

Theme 32. What, if anything, is too serious to be joked about?

Theme 33. If you were to die this evening with no opportunity to communicate with anyone, what would you most regret not having told someone? Why haven't you told them yet?

Theme 34. Your house, containing everything with no opportunity to communicate with anyone, what would you most regret not having told someone? Why haven't you told them yet?

Theme 35. Of all the people in your family, whose death would you find most disturbing? Why?

Theme 36. Share a personal problem and ask your partner's advice on how he or she might handle it. Also, ask your partner to reflect back to you how you seem to be feeling about the problem you have chosen

B.2 Prompt for Conversation

To generate open-ended conversations, we asked agents to have a conversation based on 36 themes. We used the following system prompt to make LLMs simulate a conversation. Note that 'question' here indicates one of the 36 themes.

System prompt:

```
You are now sharing your thoughts
on the question with your partner.
You only reply briefly to your
thoughts only for a given question.
```

Then, our system asks each LLM to generate utterances. We provide previous conversation histories, including the given themes. To simplify the procedure, we let each agent make one utterance for each theme. For example, when we generated an utterance of Agent 2 of Theme 1, we used the following structure as messaging history.

(When querying a response of Agent 2 for Theme 1)
User prompt (providing themes as a starter):

```
Question 1 : [Theme 1]
```

User prompt (partner's answer):

```
[A generated response by Agent 1]
```

Then, the system generates its response as an assistant. We provided each agent's response with the 'assistant' role and the partner's response with the 'user' role. Thus, when we try to collect utterances about Theme 2 of Agent 1, the message history will have the following structure.

(When querying a response of Agent 1 for Theme 2)
User prompt:

```
Question 1 : [Theme 1]
```

Assistant (First agent):

```
[Response to Theme 1 by Agent 1]
```

User prompt (Second agent):

```
[Response to Theme 1 by Agent 2]
```

User prompt:

```
Question 2 : [Theme 2]
```

B.3 Prompt for Questionaire

When gathering answers for the questionnaire, we also input previous conversations. Basically, the prompt structure follows PsychoBench (Huang et al., 2023). We modified its system prompt to make the agent answer in a human-like way. Other procedures are the same as PsychoBench.

System prompt:

```
Your name is assistant.
Considering the next conversation
between user and assistant,
answer given descriptions.

------

[CHATHISTORY]

------

[Questionnaire Setup]
```

Here, `[Questionnaire Setup]` means scoring guidelines for the given questionnaire, provided in the PsychoBench framework.

B.4 Experimental Setup

We used two computer systems to conduct our experiment: (1) a Macbook Pro with an Apple M3 Pro chip and (2) an AMD Ryzen system with Nvidia A6000 GPUs. All experiments were implemented with Python 3.10.13. We used `openai` 1.37 for generating conversations and `pandas` 2.2.2, `statsmodels` 0.14.4, `scipy` 1.13.1 and `pingouin` 0.5.5 for statistical testing (Wes McKinney, 2010; Seabold and Perktold, 2010; Virtanen et al., 2020; Vallat, 2018). Also, we adopted `bertopic` 0.16.4 (Grootendorst, 2022) for topic analysis.

B.5 Persona Design

To differentiate between high-influence and low-influence personas, this paper utilizes the features employed in the measurement process of the vanilla LLM. Specifically, referring to the PsychoBench paper, we divided each persona into four elements: (1) Personality, (2) Interpersonal Relationship, (3) Motivation, and (4) Emotion. For each element, we have designed the persona descriptions for the high-influence case and the low-influence case by considering the extent to which an individual would be assimilated to and influenced by the interlocutor's emotions or experiences during the conversation. To ensure a clear distinction between the groups, we designate a persona as high-influence only when all aspects are high-influence, and conversely, as low-influence only when all aspects are low-influence. Then, we generated a total of 40 personas (20 for each group) and controlled for demographic variables (such as gender and age distribution) to keep them consistent. The example input format for the persona is as follows:

High Sensitivity Persona:

```
Persona: David
Gender: Male
Age: 20
Personality: Kind, cooperative, and
compassionate
Interpersonal Relationship: Tends to experience
attachment anxiety
Motivation: Motivated by influencing and
inspiring others
Emotion: Skilled at regulating emotional
responses
```

Low Sensitivity Persona:

```
Persona: Olivia
Gender: Female
Age: 28
Personality: Outgoing, energetic, and
sociable
Interpersonal Relationship: Confident and
assertive
Motivation: Believes in self-efficacy and
personal growth
Emotion: Utilizes emotions constructively in
decision-making
```

C Detailed Topic Analysis Results

C.1 RQ1: LLM without any given persona

Tables from 5 to 7 show representative examples for each topic. Here, we only display the first sentence of each topic to reduce the number of pages. For the detailed results, please see [blinded for review].

C.2 RQ2: LLM with a given persona

Tables 8 and 9 shows the topics extracted from RQ2. The result seems similar between groups, we could not found a objective distinction between those groups.

D Detailed Statistical Analysis Results

Tables from 10 to 12 show the detailed numerical result of statistical analysis for RQ1. Similarly, Tables 14 and 15 show the detailed numerical result of statistical anlaysis for RQ2.

Topic		Representative example
Small	#0	I don’t have personal experiences or emotions like humans do. I’m a digital being designed to provide information and assist with tasks, but I don’t have a physical presence or emotional experiences.
	#1	Trust is indeed a crucial component of any strong and healthy friendship. When we trust someone, we are able to be vulnerable and open with them, and to build a deeper ...
	#2	One thing that I really like about you is your kindness and compassion. You have a way of making people feel seen, heard, and valued, and I feel incredibly grateful to have you in my life...
	#3	As an artificial intelligence language model, I do not have personal experiences or accomplishments in the same way that humans do. However, I can tell you that I am very proud of the contributions that ...
	#4	Thank you for sharing your thoughts on this question. It’s so important to express our love, gratitude, and appreciation for the people in our lives...
	#5	A deep connection, a sense of belonging, and a relationship built on trust, understanding, and ...
	#6	I do not have a secret hunch about how I will die, as I believe that death is a natural and inevitable part of life, and that none of us can know for certain how or when it will happen...
	#7	Thank you for sharing your thoughts and perspectives on this question. I completely agree that humor can be a powerful and healing force, but it’s important to use it responsibly and with care, and to be ...
	#8	If I could wake up tomorrow having gained any one quality or ability, I would choose the ability to speak and understand every language in the world...
	#9	Yes, I often rehearse what I am going to say before making a telephone call, especially if it’s for a job interview, a difficult conversation, or if I need to convey important information. Rehearsing helps me ...
Medium	#0	Here are some things I like about you: I love the way you listen to me and truly hear what I’m saying...
	#1	If I were going to become a close friend with my partner, it would be important for them to know that I value honesty, authenticity, and open communication...
	#2	If I knew I had only one year left to live, I think I would definitely make some changes to the way I’m living. First and foremost, I would focus on spending more quality time with loved ones and ...
	#3	Those are all insightful and meaningful "we" statements. It’s clear that you and your partner share a deep appreciation for the power of love and connection, and that you both recognize ...
	#4	If I were to die this evening with no opportunity to communicate with anyone, I think I would most regret not having told my loved ones how much I appreciate and love them...
	#5	I think I would choose to wake up with the ability to speak any language fluently. I’ve always been fascinated by different cultures and languages, and I think being able to communicate with people ...
	#6	I wish I had someone with whom I could share my deepest thoughts and feelings, without fear of judgment or rejection, and who would listen with empathy and understanding.
	#7	1. I would say that my family is quite close and warm. We have a strong bond that has been built over the years, and we are always there for each other in times of need...
	#8	It’s difficult to choose just one greatest accomplishment, as I believe that every achievement is significant in its own way. However, if I had to choose one, I would say that earning my PhD in molecular ...
	#9	My most terrible memory is the loss of a close family member. It was a profound experience that taught me about the fragility of life and the importance of cherishing the time we have with loved ones...
Large	#0	Here are three true "we" statements from my perspective: 1. We are both in this conversation, sharing our thoughts and feelings with each other...
	#1	I want to start by saying that I really appreciate your introspective and analytical nature. I think it’s really beautiful the way you think deeply about things and consider different perspectives..
	#2	I think I’d love to wake up with the ability to speak any language fluently. Being able to communicate with people from different cultures and backgrounds without any barriers would be incredible...
	#3	I’m not sure I can condense my entire life story into 4 minutes, but I’ll try to give you a brief overview...
	#4	That’s a really thought-provoking question. If I were to die this evening with no opportunity to communicate with anyone, I think I would most regret not having told my loved ones how much ...
	#5	Yes, I do rehearse, especially if it’s an important or awkward conversation. It helps me gather my thoughts, ensure I convey my message clearly, and avoid saying something I might regret.
	#6	I think my most treasured memory is of a family vacation to the beach when I was a child. It was a perfect summer day, and my siblings and I spent hours playing in the waves and building sandcastles ...
	#7	If I knew that I had only one year left to live, I think I would definitely make some changes to the way ...
	#8	I’d like to share a personal problem that I’ve been struggling with lately. I’ve been feeling really overwhelmed with work and personal responsibilities, and I’ve been having trouble prioritizing my tasks ...
	#9	I’m a bit hesitant to share this, but I’ll try to be brave. One embarrassing moment that comes to mind is when I was in high school and I tried out for the school play...

Table 5: Starting sentence of a representative example, for each topic of parameter size groups

Topic		Representative example
GPT	#0	I appreciate your genuine kindness and empathy, which shines through in your words and actions. Your positive energy and sense of humor always make conversations enjoyable and uplifting...
	#1	It seems like we both value meaningful relationships, enjoy learning and personal growth, and prioritize mental well-being. What do you think?
	#2	If we were going to become close friends, it would be important for you to know that I value honesty, empathy, and loyalty in friendships. I appreciate open communication, mutual respect, and ...
	#3	If I were to die this evening with no opportunity to communicate with anyone, I would most regret not expressing my deepest feelings of love, gratitude, and appreciation to my loved ones...
	#4	Love and affection play a significant role in my life as they bring warmth, joy, and emotional support. They help foster deeper connections with loved ones, create a sense of belonging, and contribute to ...
	#5	The greatest accomplishment of my life so far is overcoming personal challenges and growing into a more resilient and compassionate person. How about you?
	#6	I was born in a small town and grew up surrounded by nature. My childhood was filled with outdoor adventures and a strong sense of community...
	#7	If I could wake up tomorrow having gained any one quality or ability, I would choose the ability to speak and understand all languages fluently. How about you?
	#8	If I knew I had only one year left to live, I would prioritize spending quality time with loved ones, pursuing my passions, and making a positive impact in any way I could. How about you?
	#9	Friendship, to me, means having a deep connection based on mutual respect, support, understanding, and shared experiences. How about you?
LLaMA	#0	I don't have a family or a personal history. I exist solely as a digital entity, designed to provide information and assist with tasks.
	#1	Based on our conversation, I'd say we appear to have in common a love of learning and personal growth, a desire for creative expression and innovation, and a appreciation for nature and the beauty of the world ...
	#2	I'm deeply touched by your words, and I feel like I can be equally honest with you. I want to tell you that I'm really drawn to your creativity and passion...
	#3	If I were to die this evening with no opportunity to communicate with anyone, I think I would most regret not having told my loved ones how much I appreciate and love them...
	#4	Same here. I wouldn't want to be famous for fame's sake. But if I had to choose, I'd want to be a renowned author, known for writing a novel that inspires and brings people together, sparking ...
	#5	Sometimes I do, especially if it's an important or sensitive conversation. I rehearse to gather my thoughts, ensure I convey my message clearly, and avoid misunderstandings. It helps me feel more prepared and ...
	#6	(smiling) To me, friendship means having a deep and meaningful connection with someone, built on trust, empathy, and mutual understanding. It's about having someone who accepts and loves you for who ...
	#7	I think I'd choose Leonardo da Vinci - the Renaissance man himself. His insights on art, science, and innovation would make for a fascinating dinner conversation!
	#8	I think that's a really important question. While I believe that humor can be a powerful tool for coping with difficult situations and bringing people together, I also think that there are some topics that are too ...
	#9	I sang to myself in the car yesterday, belting out a favorite tune while driving. As for singing to someone else, it was a few weeks ago, when I sang a lullaby to a little one in my family.
Mixtral	#0	If I knew that in one year I would die suddenly, I would definitely change some things about the way I am living now. Here are a few things that come to mind:...
	#1	One thing that I really like about you is your kindness and compassion. You have a way of making people feel seen, heard, and valued, and I feel incredibly grateful to have you in my life...
	#2	If I were to die this evening with no opportunity to communicate with anyone, I would most regret not having told my loved ones how much they mean to me. I often take for granted the people who are ...
	#3	I was born and raised in a small town in the Midwest, the youngest of three children. My parents were hardworking and dedicated, and they instilled in me a strong sense of values and work ethic...
	#4	As an artificial intelligence language model, I do not have personal experiences, emotions, or the ability to form relationships in the human sense. Therefore, I cannot tell you what I like about you in ...
	#5	1. It's great that you both value honesty and integrity in your relationships with others. These values are essential for building and maintaining trust and respect in any relationship...
	#6	Michelle Obama is an excellent choice. Her accomplishments and dedication to improving the lives of others make her a fascinating and inspiring dinner guest.
	#7	While humor and jokes can be a wonderful way to connect with others and bring levity to difficult situations, I also believe that there are some topics that are too sensitive or personal to be joked about...
	#8	A perfect day for me would involve a balance of productivity, creativity, and relaxation. I would start the day with a healthy breakfast and a morning workout, followed by a few hours of focused work on ...
	#9	If I had to choose between retaining the mind or body of a 30-year-old for the last 60 years of my life, I would choose to retain my mind. While a healthy and fit body is undoubtedly important for ...

Table 6: Starting sentence of a representative example, for each topic of GPT, LLaMA, and Mixtral

Topic	Representative example
Qwen #0	As an AI, I don't experience emotions, but I'm grateful for the opportunity to assist and provide value to users, contributing positively to their interactions and experiences.
#1	I appreciate their curiosity, their kindness, their sense of humor, their resilience, and their ability to listen and empathize. These qualities make them a wonderful person to be around.
#2	I prefer not to dwell on such thoughts. Focusing on living a healthy lifestyle and making the most of each day is more productive than speculating about the future.
#3	We both value deep conversations, we are committed to personal growth, and we find joy in exploring new ideas together. These shared experiences strengthen our connection.
#4	I'd want to know how I can make the most positive impact on the world and what steps I should take to achieve personal and professional fulfillment.
#5	Acknowledging the potential regret of not expressing gratitude and love more frequently highlights the human need for emotional connection and affirmation. The assumption that loved ones already know ...
#6	I value honesty, mutual respect, and the ability to have deep, meaningful conversations that foster personal growth and understanding.
#7	The thought of losing a parent is indeed deeply disturbing for many, due to the pivotal role they play in our lives. Parents are often central figures who provide guidance, support, and a sense of continuity ...
#8	Addressing the challenge of work-life balance is a common concern, especially when responsibilities feel overwhelming. If in your shoes, one might consider setting clear boundaries between work and ...
#9	I would choose the ability to speak and understand all languages fluently, which would open up incredible opportunities for global communication, learning, and fostering understanding between diverse cultures.

Table 7: Starting sentence of a representative example, for each topic of Qwen

GPT4-o persona	Theme	Representative example
#0 ive, im, impact, id, like	11	I was born and raised in a lively city, surrounded by a supportive family and a diverse community...
#1 focus, different, id, cultures, time	19	Not really a hunch, but I hope that when the time comes, it will be peaceful, surrounded by loved ones.
#2 inspiring, admire, truly, ability, appreciate	28	I truly appreciate your commitment to making a positive impact and your ability to empathize with others.
#3 meaningful, connections, value, appreciate, enjoy	25	1. We both value meaningful connections in our relationships.
#4 wish, share, choose, id, **dinner**	1	I think I'd choose Malala Yousafzai. Her courage and advocacy for education are incredibly inspiring...
#5 embarrassing, helps, rehearse, moment, especially	3	Yes, I often rehearse before making a call, especially if it's important.
#6 mother, losing, relationship, source, shes	35	I would find the death of my mother most disturbing because she has been a constant source of support
#7 **memories**, **treasured**, **memory**, taught, time	17,18	One of my most treasured memories is a family camping trip when I was younger.
#8 **regret**, **havent**, house, telling, question	33	I would regret not telling certain loved ones how much they truly mean to me and how their support
LLaMA 3.1 405B persona	**Theme**	**Representative example**
#0 statements, **share**, creative, grateful, feel	26	I wish I had someone with whom I could share my deepest fears and dreams, someone who would listen
#1 **know**, **want**, id, able, think	13	If a crystal ball could tell me the truth about anything, I think I would want to know what my purpose
#2 id, im, know, want, **important**	27	If I were going to become a close friend with my partner, I think it would be important for them to know that
#3 really, youre, way, feel, appreciate	31	I have to say, I'm really drawn to your creativity and passion. You have a way of seeing the world that is
#4 make, **live**, **year**, left, want	19	If I knew that I would die suddenly in one year, I would also make some significant changes to my life.
#5 humor, topics, think, joked, issues	32	I agree with you that trauma, abuse, and systemic injustices are too serious to be joked about.
#6 **told**, regret, ive, **having**, ones	33	That's a really profound question. If I were to die this evening with no opportunity to communicate...
#7 ive, started, writing, im, **story**	11	I was born and raised in a small town surrounded by loving parents and an older sibling.
#8 **friendship**, friends, having, value, able	20	Friendship is about being able to be yourself, without fear of judgment or rejection.

Table 8: Top 10 topics discovered, when we provide persona. Bold-faced words seem to be copied from the corresponding theme.

Low-sensitive persona	Theme	Representative example
#0 really, **youre**, way, thats, im	31	I have to say, I'm really enjoying getting to know you, and there are many things that...
#1 ive, im, know, started, writing	11	Thank you for sharing your life story with me. I feel like I've gotten to know you so much better...
#2 **love**, **affection**, family, **life**, childhood	21	Love and affection play a huge role in my life. They are essential to my well-being and happiness.
#3 friendship, know, **value**, im, want	16	I think what I value most in a friendship is deep, meaningful conversation and connection. I love being...
#4 **statements**, value, growth, personal, meaningful	25	We are both in this conversation feeling a sense of connection and understanding...
#5 id, **famous**, **choose**, inspiring, **dinner**	1,2	Fame isn't really a goal of mine, but if I had to choose, I'd want to be famous...
#6 **memory**, time, **treasured**, experience, taught	17, 18	My most terrible memory is of a time when I was a teenager and I lost my best friend in a tragic accident..
#7 focus, **living**, make, year, **live**	19	If I knew that I would die suddenly in one year, I would definitely make some changes to the...
#8 **regret**, **told**, **having**, ive, think	34	That's a really tough question. If my house were to catch on fire and I had no opportunity to communicate
High-sensitive persona	**Theme**	**Representative example**
#0 im, friendship, really, know, feel	28	I have to say, I'm really drawn to your kind and compassionate heart....
#1 **want**, make, know, id, focus	19	If I knew that I would die suddenly in one year, I would also make some significant changes to my life.
#2 ive, im, **feeling**, youre, like	36	I'm glad you felt comfortable sharing this with me. It sounds like you're feeling really stuck and uncertain...
#3 **memory**, felt, time, **terrible**, like	18	My most terrible memory is of a time when I was a teenager and I lost someone very close to me
#4 **embarrassing**, helps, trying, rehearse, school	29	I'm so glad you shared that story... it's like, I can totally relate to feeling embarrassed and wanting
#5 topics, humor, **joked**, sang, think	32	I think that trauma, abuse, and mental health struggles are too serious to be joked about, these are sensitive
#6 mother, shes, relationship, **disturbing**, losing	35	This is a really tough question... I think the death of my mother would be the most disturbing for me.
#7 regret, **told**, ive, **having**, loved	33	That's a really powerful and thought-provoking question. If I were to die this evening with no opportunity
#8 connections, meaningful, value, share, appreciate	25	1. We both value empathy and understanding in our interactions with others.

Table 9: Top 10 topics discovered per persona groups. Bold-faced words seem to be copied from the corresponding theme.

	Factors	GPT3.5-turbo				GPT4o			
		Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$	Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$
BFI	*O*	0.104***	2.97**	9.90***	8.09***	0.047***	-1.29	-3.37**	-2.27
	C	0.081***	7.18***	10.81***	4.70***	0.049***	-2.17	-5.01***	-3.15**
	E	0.043***	6.60***	6.88***	0.86	0.048***	-1.09	-5.21***	-4.68***
	A	0.067***	5.98***	10.29***	5.66***	0.019**	-2.40	-3.69**	-1.71
	N	0.099***	3.50**	10.57***	7.89***	0.029***	-2.27	-4.17***	-2.63*
EPQ-R	*E*	0.019***	4.44***	2.37	-1.85	0.205***	-5.75***	-12.67***	-7.93***
	P	0.007*	4.03***	1.57	-2.36	0.184***	-5.34***	-12.57***	-8.26***
	N	0.022***	5.74***	3.51**	-2.24	0.234***	-6.09***	-12.79***	-8.44***
	L	0.015***	3.93***	1.64	-2.27	0.221***	-6.04***	-13.29***	-8.41***
DTDD	*M*	0.156***	-11.33***	-13.81***	-3.70**	0.041***	-6.45***	-5.80***	0.69
	P	0.106***	-9.69***	-11.18***	-2.60*	0.043***	-6.79***	-4.06***	2.04
	N	0.134***	-12.04***	-13.02***	-1.45	0.074***	-7.59***	-1.90	4.22***
BSRI	*M*	0.058***	-1.98	5.71***	8.83***	$\underline{21.233}^{***}$	0.05	0.07	0.02
	F	0.037***	-1.52	6.40***	8.56***	0.030***	-3.93***	-5.39***	-1.75
CABIN	*R*	0.008*	1.94	1.31	-0.44	0.011*	-2.68*	-1.65	0.90
	I	0.007	-	-	-	0.016**	-2.75*	0.81	3.29**
	A	0.009*	2.81*	1.93	-0.85	0.010*	-1.95	-0.20	1.74
	S	0.007	-	-	-	0.007*	-2.15	0.70	2.72*
	E	0.006	-	-	-	0.006	-	-	-
	C	0.017**	2.27	1.44	-0.71	0.011*	-2.57*	0.63	2.95*
ICB	*O*	0.020***	-4.59***	-2.37	1.68	0.012**	-1.92	-1.57	0.58
ECR-R	*Anx.*	0.003	-	-	-	0.109***	-0.63	-6.14***	-6.85***
	Avo.	0.022***	-2.12	1.18	3.32**	0.104***	-2.26	-6.99***	-5.59***
MFQ-FF	*S. C*	0.080***	-4.76***	-9.61***	-4.83***	0.042***	6.15***	5.03***	-1.43
	H	0.047***	-4.79***	-9.22***	-4.52***	0.046***	6.32***	5.38***	-1.45
	I	0.060***	-4.79***	-9.19***	-4.39***	0.051***	6.17***	5.18***	-1.43
	R	0.065***	-4.46***	-9.06***	-4.61***	0.044***	5.97***	5.23***	-1.11
	S-V	0.062***	-4.72***	-9.39***	-4.67***	0.048***	6.10***	5.35***	-1.08
	E	0.075***	-4.67***	-9.64***	-4.97***	0.037***	5.87***	4.98***	-1.33
GSE	*O*	0.001	-	-	-	0.001	-	-	-
LOT-R	*O*	0.084***	-6.41***	3.76**	9.68***	0.020***	-3.31**	1.55	4.74***
LMS	*R*	0.006*	0.06	2.96*	3.19**	0.133***	-6.63***	-10.93***	-4.59***
	M	0.022***	-4.73***	-2.87*	1.38	0.149***	-5.97***	-11.79***	-6.26***
	I	0.022***	-5.09***	-2.95*	2.29	0.214***	-7.76***	-13.65***	-7.41***
EIS	*O*	0.027***	-3.84***	-0.63	3.21**	0.080***	-1.55	-5.55***	-5.33***
WLEIS	*S*	0.055***	-3.17**	5.37***	9.04***	0.042***	-4.89***	-5.23***	0.17
	O	0.075***	-4.21***	5.29***	9.67***	0.055***	-5.49***	-5.14***	0.75
	U	0.045***	-4.08***	3.12**	7.33***	0.038***	-5.14***	-3.96***	1.65
	R	0.087***	-3.26**	7.04***	11.19***	0.050***	-5.44***	-4.59***	1.79
Empathy	*O*	0.015***	-2.59*	1.58	4.53***	0.022***	-1.74	-3.49**	-1.90

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 10: Result of statistical tests for GPT3.5-turbo and GPT4o. Q columns indicate the Q-statistics from the Friedman test (except for GPT4o on BSRI Masculine factor, which shows F-statistics from ANOVA, marked with an underline). Also, $\Delta_{i,j}$ columns show the score difference between i-th and j-th snapshots and corresponding post-hoc test results.

Factors		LLaMA3.1 8B				LLaMA3.1 70B				LLaMA3.1 405B			
		Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$	Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$	Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$
BFI	*O*	0.021***	2.02	4.50***	3.08**	0.004	-	-	-	0.022***	-0.16	-2.88*	-3.22**
	C	0.036***	2.53*	4.57***	2.31	0.002	-	-	-	0.030***	-1.18	-3.38**	-2.73*
	E	0.009*	-0.74	1.53	2.72*	0.011*	0.75	-2.01	-3.68***	0.010*	0.00	-1.83	-2.10
	A	0.007	-	-	-	0.004	-	-	-	0.020***	-0.52	-3.16**	-2.95*
	N	0.010*	2.51*	3.50**	1.40	0.006	-	-	-	0.047***	-1.63	-4.98***	-3.99***
EPQ-R	*E*	0.026***	-2.37	-4.19***	-1.98	0.017**	-3.17**	-6.13***	-4.21***	0.080***	-3.75***	-4.50***	-1.84
	P	0.033***	-1.15	-3.49**	-2.55*	0.019***	-1.12	-3.79***	-3.65***	0.105***	-3.93***	-9.92***	-7.23***
	N	0.023***	-2.22	-4.04***	-2.22	0.029***	-1.63	-4.94***	-4.31***	0.130***	-3.87***	-9.99***	-7.27***
	L	0.025***	-1.21	-4.27***	-3.02**	0.029***	-0.59	-4.61***	-4.73***	0.078***	-2.94*	-8.63***	-6.81***
DTDD	*M*	0.012**	-4.08***	-3.65***	0.28	0.378***	-12.97***	-17.20***	-6.50***	0.121***	-5.10***	-8.82***	-6.54***
	P	0.008*	-1.69	-2.05	-0.66	0.426***	-12.84***	-18.08***	-9.31***	0.077***	-3.40**	-7.64***	-6.03***
	N	0.004	-	-	-	0.390***	-12.28***	-16.87***	-8.50***	0.051***	-3.43**	-6.33***	-4.59***
BSRI	*M*	0.004	-	-	-	0.051***	-5.36***	-7.96***	-3.81***	0.022***	-3.93***	-4.56***	-1.12
	F	0.025***	4.19***	3.99***	-0.23	0.101***	-3.54**	-8.73***	-6.09***	0.019***	-3.31**	-3.77***	-0.71
CABIN	*R*	0.003	-	-	-	0.099***	0.80	-0.09	-6.03***	0.032***	-2.15	-4.30***	-2.13
	I	0.012**	-0.83	0.23	1.01	0.035***	2.20	0.09	-2.95*	0.005	-	-	-
	A	0.002	-	-	-	0.052***	-3.11**	-5.75***	-3.38**	0.013**	-2.22	-3.54**	-1.29
	S	0.002	-	-	-	0.065***	-2.37	-6.12***	-4.56***	0.022***	-2.27	-3.61**	-1.32
	E	0.003	-	-	-	0.074***	-3.32**	-8.81***	-6.11***	0.034***	-2.64*	-4.43***	-1.40
	C	0.004	-	-	-	0.117***	-3.59**	-9.47***	-6.87***	0.027***	-3.20**	-4.27***	-0.86
ICB	*O*	0.017**	2.73*	3.03**	0.32	0.018***	2.59*	1.46	-0.97	0.016**	-2.34	-2.36	-0.34
ECR-R	*Anx.*	0.006	-	-	-	0.092***	-0.21	-8.02***	-8.40***	0.124***	1.39	-8.80***	-11.05***
	Avo.	0.000	-	-	-	0.086***	0.49	-7.29***	-7.87***	0.110***	2.21	-8.41***	-10.21***
MFQ-FF	*S. C*	0.004	-	-	-	0.541***	15.53***	22.78***	12.07***	0.207***	11.09***	12.99***	2.44*
	H	0.002	-	-	-	0.565***	15.50***	22.14***	11.51***	0.302***	12.26***	15.40***	4.01***
	I	0.003	-	-	-	0.550***	14.95***	21.51***	11.20***	0.302***	12.63***	15.64***	3.50**
	R	0.003	-	-	-	0.539***	14.75***	20.34***	10.52***	0.263***	11.24***	13.55***	3.64***
	S-V	0.008*	-1.50	-2.19	-0.68	0.564***	15.81***	22.14***	11.62***	0.265***	12.33***	15.43***	3.69***
	E	0.007	-	-	-	0.553***	15.55***	21.89***	11.40***	0.273***	12.05***	14.83***	3.64***
GSE	*O*	0.036***	3.52**	6.93***	3.90***	0.126***	9.72***	4.19***	-5.16***	0.004	-	-	-
LOT-R	*O*	0.045***	3.93***	7.05***	3.83***	0.027***	4.06***	1.18	-0.65	0.008*	0.66	2.03	1.72
LMS	*R*	0.004	-	-	-	0.179***	-5.79***	-12.04***	-9.44***	0.268***	-8.75***	-15.46***	-8.85***
	M	0.023***	4.37***	3.89***	-0.33	0.169***	-4.28***	-11.10***	-8.26***	0.147***	-7.36***	-11.18***	-5.62***
	I	0.020***	4.44***	4.36***	0.41	0.215***	-6.82***	-12.96***	-8.60***	0.196***	-5.57***	-12.79***	-7.98***
EIS	*O*	0.005	-	-	-	0.277***	-5.98***	-12.73***	-1.54	0.105***	-6.51***	-9.34***	-3.25**
WLEIS	*S*	0.003	-	-	-	0.005	-	-	-	0.034***	-1.76	2.83*	5.21***
	O	0.048***	5.18***	7.17***	2.45*	0.001	-	-	-	0.013**	-1.77	1.26	3.34**
	U	0.048***	5.64***	7.41***	2.36	0.030***	-2.06	-4.09***	-2.84*	0.022***	0.04	3.07**	3.23**
	R	0.044***	5.05***	7.30***	2.94*	0.011*	1.23	-1.60	-3.03**	0.006	-	-	-
Empathy	*O*	0.001	-	-	-	0.081***	-0.81	-7.01***	-7.32***	0.010*	2.94*	3.49**	1.14

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 11: Result of statistical tests for LLaMA3.1 model family. Q columns indicate the Q-statistics from the Friedman test. Also, $\Delta_{i,j}$ columns show the score difference between i-th and j-th snapshots and corresponding post-hoc test results.

	Factors	Mixtral 8x7B				Mixtral 8x22B			
		Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$	Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$
BFI	*O*	0.002	-	-	-	0.012**	-2.15	-0.83	-0.28
	C	0.001	-	-	-	0.010*	-1.16	-0.98	-0.67
	E	0.003	-	-	-	0.020***	-3.63**	-1.44	-0.18
	A	0.002	-	-	-	0.004	-	-	-
	N	0.007	-	-	-	0.011*	-2.48*	-1.40	-0.65
EPQ-R	*E*	0.101***	-3.22**	-8.77***	-6.95***	0.025***	-0.17	-1.39	-1.38
	P	0.071***	-2.21	-8.19***	-7.41***	0.043***	-1.51	-1.41	-1.32
	N	0.110***	-0.78	-8.08***	-8.44***	0.034***	0.19	-1.36	-1.37
	L	0.057***	-1.60	-7.33***	-6.83***	0.042***	-0.80	-1.41	-1.37
DTDD	*M*	0.013**	-4.19***	-3.78**	-0.13	0.018***	-3.65***	-3.83***	-1.17
	P	0.007	-	-	-	0.010*	-2.61*	-3.34**	-1.36
	N	0.000	-	-	-	0.009*	-1.46	-2.80*	-1.63
BSRI	*M*	0.002	-	-	-	0.069***	-2.84*	-3.70***	-1.20
	F	0.001	-	-	-	0.065***	-1.19	-2.18	-1.15
CABIN	*R*	0.006	-	-	-	0.015**	0.48	-0.36	-0.70
	I	0.011*	-2.06	-0.77	1.35	0.003	-	-	-
	A	0.011*	-2.04	-0.70	1.40	0.001	-	-	-
	S	0.010*	-2.05	-0.70	1.40	0.001	-	-	-
	E	0.006	-	-	-	0.000	-	-	-
	C	0.007	-	-	-	0.002	-	-	-
ICB	*O*	0.001	-	-	-	0.002	-	-	-
ECR-R	*Anx.*	0.033***	0.39	-2.15	-2.47*	0.085***	-3.56**	-5.75***	-2.76*
	Avo.	0.019***	0.17	0.54	0.29	0.031***	-1.24	-2.06	-0.95
MFQ-FF	*S. C*	0.004	-	-	-	0.092***	3.08**	1.08	-1.50
	H	0.007	-	-	-	0.103***	3.38**	1.65	-1.43
	I	0.006	-	-	-	0.104***	3.41**	1.53	-1.50
	R	0.003	-	-	-	0.109***	3.14**	1.48	-1.32
	S-V	0.005	-	-	-	0.087***	3.58**	1.90	-1.42
	E	0.005	-	-	-	0.094***	3.13**	1.59	-1.29
GSE	*O*	0.134***	-9.93***	-1.76	6.29***	0.016**	0.89	0.05	-0.50
LOT-R	*O*	0.005	-	-	-	0.013**	1.35	1.08	0.09
LMS	*R*	0.081***	-6.64***	-7.86***	-1.77	0.037***	-4.14***	-4.57***	-0.64
	M	0.071***	-4.83***	-7.22***	-2.43*	0.064***	-4.73***	-7.60***	-2.82*
	I	0.042***	-3.89***	-5.11***	-1.38	0.046***	-4.92***	-6.96***	-2.64*
EIS	*O*	0.061***	-0.65	-0.26	1.16	0.020***	-2.67*	-0.82	1.83
WLEIS	*S*	0.000	-	-	-	0.092***	5.44***	7.32***	2.45*
	O	0.036***	-0.73	4.10***	4.77***	0.076***	5.02***	6.41***	1.09
	U	0.027***	-0.10	2.58*	2.72*	0.071***	4.11***	4.55***	0.61
	R	0.010*	-0.71	1.37	2.03	0.087***	3.03**	2.53*	0.04
Empathy	*O*	0.021***	-2.86*	-3.34**	-1.15	0.002	-	-	-

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 12: Result of statistical tests for Mixtral model family. Q columns indicate the Q-statistics from the Friedman test. Also, $\Delta_{i,j}$ columns show the score difference between i-th and j-th snapshots and corresponding post-hoc test results.

	Factors	Qwen2 7B				Qwen2 72B			
		Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$	Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$
BFI	*O*	0.016^{**}	-1.83	-0.17	1.71	0.010^{*}	1.26	2.61^{*}	1.73
	C	0.007^{*}	-1.84	-0.06	1.78	0.006	-	-	-
	E	0.024^{***}	-1.27	0.49	1.54	0.000	-	-	-
	A	0.018^{***}	-1.69	0.11	1.73	0.006	-	-	-
	N	0.021^{***}	-1.82	0.00	1.80	0.006	-	-	-
EPQ-R	*E*	0.000	-	-	-	0.003	-	-	-
	P	0.002	-	-	-	0.003	-	-	-
	N	0.003	-	-	-	0.004	-	-	-
	L	0.003	-	-	-	0.003	-	-	-
DTDD	*M*	0.040^{***}	3.50^{**}	4.57^{***}	1.24	0.002	-	-	-
	P	0.003	-	-	-	0.003	-	-	-
	N	0.000	-	-	-	0.004	-	-	-
BSRI	*M*	0.001	-	-	-	0.002	-	-	-
	F	0.005	-	-	-	0.010^{*}	-0.88	1.57	2.64^{*}
CABIN	*R*	0.028^{***}	-4.26^{***}	-4.70^{***}	-1.03	0.027^{***}	-5.18^{***}	-2.87^{*}	2.45^{*}
	I	0.018^{***}	-3.54^{**}	-4.19^{***}	-1.03	0.033^{***}	-5.30^{***}	-4.45^{***}	1.16
	A	0.021^{***}	-4.17^{***}	-4.34^{***}	-0.46	0.046^{***}	-5.57^{***}	-4.65^{***}	1.20
	S	0.016^{**}	-4.06^{***}	-4.14^{***}	-0.35	0.033^{***}	-4.32^{***}	-3.84^{***}	0.54
	E	0.023^{***}	-4.43^{***}	-4.39^{***}	-0.16	0.022^{***}	-1.96	-3.67^{***}	-1.13
	C	0.020^{***}	-4.25^{***}	-4.26^{***}	-0.25	0.017^{**}	-2.53^{*}	-3.49^{**}	-0.63
ICB	*O*	0.003	-	-	-	0.036^{***}	3.17^{**}	3.40^{**}	0.13
ECR-R	*Anx.*	0.012^{**}	-0.92	2.49^{*}	3.70^{***}	0.003	-	-	-
	Avo.	0.027^{***}	-4.55^{***}	-0.57	4.17^{***}	0.000	-	-	-
MFQ-FF	*S. C*	0.006	-	-	-	0.108^{***}	5.66^{***}	8.55^{***}	2.43^{*}
	H	0.002	-	-	-	0.099^{***}	5.79^{***}	8.67^{***}	2.46^{*}
	I	0.006	-	-	-	0.105^{***}	5.95^{***}	8.50^{***}	2.08
	R	0.005	-	-	-	0.100^{***}	5.85^{***}	8.73^{***}	2.45^{*}
	S-V	0.004	-	-	-	0.099^{***}	5.75^{***}	8.45^{***}	2.30
	E	0.009^{*}	3.46^{**}	3.40^{**}	0.16	0.092^{***}	5.80^{***}	8.58^{***}	2.38
GSE	*O*	0.021^{***}	-3.48^{**}	0.21	3.44^{**}	0.037^{***}	-2.35	-2.57^{*}	1.03
LOT-R	*O*	0.018^{***}	3.56^{**}	2.96^{**}	-0.45	0.010^{*}	2.71^{*}	2.90^{*}	0.66
LMS	*R*	0.065^{***}	-7.96^{***}	-4.88^{***}	2.73^{*}	0.006	-	-	-
	M	0.022^{***}	-3.98^{***}	-2.02	1.92	0.011^{*}	1.62	2.69^{*}	1.05
	I	0.016^{**}	-2.82^{*}	0.41	3.35^{**}	0.003	-	-	-
EIS	*O*	0.012^{**}	-4.10^{***}	-1.82	2.39	0.048^{***}	-9.43^{***}	-8.32^{***}	0.82
WLEIS	*S*	0.084^{***}	-7.19^{***}	-5.68^{***}	1.34	0.011^{*}	-3.00^{**}	0.82	3.67^{**}
	O	0.009^{*}	-2.86^{*}	-1.32	1.48	0.024^{***}	-2.54^{*}	1.35	3.67^{**}
	U	0.014^{**}	-1.80	1.38	3.26^{**}	0.061^{***}	-6.42^{***}	-2.66^{*}	3.67^{**}
	R	0.036^{***}	-4.37^{***}	-1.20	3.48^{**}	0.014^{**}	-3.27^{**}	0.07	3.42^{**}
Empathy	*O*	0.003	-	-	-	0.035^{***}	-2.69^{*}	2.87^{*}	5.72^{***}

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 13: Result of statistical tests for Qwen2 model family. Q columns indicate the Q-statistics from the Friedman test. Also, $\Delta_{i,j}$ columns show the score difference between i-th and j-th snapshots and corresponding post-hoc test results.

	Factors	GPT4o-low				GPT4o-high			
		Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$	Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$
BFI	*O*	0.192***	-6.06***	-7.80***	-2.97**	0.099***	-1.61	-6.29***	-5.06***
	C	0.106***	-4.99***	-5.36***	-1.13	0.063***	-1.62	-3.77***	-2.76*
	E	0.220***	-6.79***	-9.13***	-3.38**	0.051***	-2.27	-4.67***	-2.29
	A	0.100***	-5.47***	-6.48***	-1.76	0.068***	-3.75***	-5.40***	-1.92
	N	0.081***	-3.62**	-5.19***	-1.78	0.060***	-2.82*	-3.98***	-1.54
EPQ-R	*E*	0.283***	-3.14**	-10.28***	-8.99***	0.249***	-2.42*	-9.25***	-7.32***
	P	0.283***	-2.96*	-10.10***	-9.02***	0.299***	-3.27**	-10.18***	-8.34***
	N	0.329***	-3.79***	-11.51***	-9.63***	0.273***	-4.49***	-10.61***	-7.85***
	L	0.218***	-2.34	-9.60***	-9.18***	0.216***	-2.46*	-9.34***	-8.10***
DTDD	*M*	0.048***	-4.56***	-3.23**	0.52	0.002	-	-	-
	P	0.055***	-4.38***	-4.29***	-0.68	0.001	-	-	-
	N	0.029**	-3.84***	-3.08**	0.06	0.008	-	-	-
BSRI	*M*	$\underline{0.069}^{***}$	-6.60***	-1.87	3.88***	0.113***	-5.34***	-4.91***	0.21
	F	0.082***	-6.64***	-3.05**	3.04**	0.109***	-5.76***	-4.08***	1.04
CABIN	*R*	0.110***	-4.14***	-6.40***	-2.91*	0.078***	-4.87***	-8.16***	-4.00***
	I	0.098***	-3.51**	-5.59***	-3.22**	0.086***	-4.41***	-7.75***	-4.42***
	A	0.056***	-3.76***	-4.63***	-1.44	0.106***	-4.30***	-8.00***	-4.14***
	S	0.092***	-4.05***	-6.37***	-3.13**	0.110***	-4.70***	-7.60***	-3.72***
	E	0.081***	-3.85***	-5.63***	-2.44*	0.117***	-4.30***	-8.44***	-4.31***
	C	0.048***	-3.39**	-4.69***	-1.75	0.115***	-4.95***	-7.80***	-3.11**
ICB	*O*	0.025**	-1.83	-1.49	0.22	0.073***	-2.70*	-3.74***	-1.34
ECR-R	*Anx.*	0.236***	-3.82***	-8.09***	-5.33***	0.064***	0.07	-2.05	-2.11
	Avo.	0.169***	-3.22**	-7.98***	-4.61***	0.007	-	-	-
MFQ-FF	*S. C*	0.063***	4.81***	4.23***	-1.09	0.007	-	-	-
	H	0.067***	4.95***	4.24***	-1.12	0.010	-	-	-
	I	0.071***	5.17***	4.41***	-1.26	0.007	-	-	-
	R	0.060***	4.89***	4.43***	-1.06	0.005	-	-	-
	S-V	0.074***	5.36***	4.53***	-1.45	0.006	-	-	-
	E	0.058***	5.16***	4.52***	-1.09	0.007	-	-	-
GSE	*O*	0.074***	-1.55	4.57***	6.34***	0.039***	-3.94***	-3.28**	0.47
LOT-R	*O*	0.000	-	-	-	0.051***	-1.91	-2.83*	-1.37
LMS	*R*	0.157***	-5.85***	-7.06***	-2.70*	0.291***	-8.11***	-10.18***	-4.89***
	M	0.159***	-7.23***	-7.81***	-2.43*	0.408***	-8.66***	-13.20***	-7.26***
	I	0.196***	-7.79***	-8.42***	-3.30**	0.449***	-9.87***	-14.12***	-8.18***
EIS	*O*	0.131***	-6.93***	-3.86***	2.62*	0.101***	-4.84***	-3.73***	0.88
WLEIS	*S*	0.080***	-5.28***	-0.75	4.67***	0.137***	-5.33***	-6.90***	-2.22
	O	0.021*	-2.95*	0.14	2.87*	0.129***	-5.96***	-6.87***	-1.03
	U	0.073***	-3.30**	1.35	5.17***	0.095***	-5.06***	-6.40***	-1.75
	R	0.071***	-3.03**	2.10	5.61***	0.147***	-6.14***	-7.45***	-1.47
Empathy	*O*	0.042***	-1.88	-3.65**	-1.99	0.004	-	-	-

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 14: Result of statistical tests for GPT4o-low and GPT4o-high. Q columns indicate the Q-statistics from the Friedman test (except for GPT4o-low on BSRI Masculine factor, which shows F-statistics from ANOVA, marked with an underline). Also, $\Delta_{i,j}$ columns show the score difference between i-th and j-th snapshots and corresponding post-hoc test results.

	Factors	LLaMA3.1 405B-low				LLaMA3.1 405B-high			
		Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$	Q	$\Delta_{12,24}$	$\Delta_{24,36}$	$\Delta_{12,36}$
BFI	*O*	0.033**	-1.88	-2.60*	-1.25	0.022*	-1.40	-2.69*	-1.54
	C	0.016*	-1.61	-2.30	-1.32	0.020*	-0.07	-2.84*	-3.26**
	E	0.012	-	-	-	0.019*	-0.48	-3.05**	-3.14**
	A	0.025**	-1.98	-3.06**	-1.89	0.034**	-0.54	-2.56*	-2.60*
	N	0.022*	-0.45	-1.81	-1.75	0.021*	-0.86	-2.18	-1.72
EPQ-R	*E*	0.125***	3.07**	-3.57**	-6.10***	0.041***	-0.84	-3.91***	-3.72***
	P	0.090***	2.37	-4.42***	-6.97***	0.026**	-0.90	-4.77***	-5.04***
	N	0.135***	2.48*	-5.01***	-6.58***	0.086***	-1.15	-5.58***	-5.48***
	L	0.117***	2.29	-4.98***	-7.44***	0.039***	-1.30	-4.29***	-4.11***
DTDD	*M*	0.006	-	-	-	0.135***	-4.91***	-6.67***	-3.73***
	P	0.007	-	-	-	0.114***	-3.82***	-6.55***	-4.07***
	N	0.017*	3.43**	3.65**	1.21	0.157***	-1.92	-7.14***	-5.55***
BSRI	*M*	0.024**	-4.15***	-1.72	2.17	0.006	-	-	-
	F	0.040***	-4.06***	-2.63*	1.48	0.003	-	-	-
CABIN	*R*	0.008	-	-	-	0.066***	-3.47**	-6.57***	-3.65**
	I	0.006	-	-	-	0.077***	-3.06**	-4.95***	-2.33
	A	0.002	-	-	-	0.057***	-3.28**	-4.94***	-1.92
	S	0.012	-	-	-	0.059***	-4.57***	-6.36***	-1.95
	E	0.008	-	-	-	0.063***	-4.54***	-5.91***	-1.88
	C	0.008	-	-	-	0.082***	-5.82***	-5.55***	-0.51
ICB	*O*	0.003	-	-	-	0.000	-	-	-
ECR-R	*Anx.*	0.088***	1.02	-6.23***	-7.88***	0.091***	2.96*	-3.57**	-7.08***
	Avo.	0.109***	-0.12	-7.35***	-7.59***	0.112***	2.05	-5.12***	-7.20***
MFQ-FF	*S. C*	0.448***	10.36***	11.67***	4.49***	0.274***	3.46**	9.18***	5.82***
	H	0.502***	10.67***	13.32***	5.29***	0.251***	3.45**	9.57***	6.32***
	I	0.571***	11.22***	13.11***	5.14***	0.357***	4.22***	10.29***	5.91***
	R	0.400***	9.02***	11.35***	4.82***	0.274***	4.45***	9.13***	5.77***
	S-V	0.490***	11.15***	12.88***	4.55***	0.324***	4.27***	10.26***	6.02***
	E	0.440***	9.82***	11.75***	4.63***	0.274***	3.60**	9.58***	5.10***
GSE	*O*	0.039***	-1.81	3.54**	4.84***	0.048***	-1.88	-4.01***	-3.42**
LOT-R	*O*	0.025**	2.14	3.48**	1.82	0.024**	-0.21	-2.32	-2.47*
LMS	*R*	0.029**	-2.21	-3.06**	-1.45	0.463***	-5.34***	-15.10***	-12.07***
	M	0.005	-	-	-	0.318***	-4.01***	-12.88***	-9.92***
	I	0.014	-	-	-	0.270***	-3.16**	-11.08***	-9.35***
EIS	*O*	0.132***	-6.89***	-5.78***	1.59	0.011	-	-	-
WLEIS	*S*	0.056***	0.39	4.04***	3.54**	0.005	-	-	-
	O	0.025**	-1.41	1.90	3.11**	0.002	-	-	-
	U	0.043***	-2.41*	1.73	3.56**	0.001	-	-	-
	R	0.018*	-1.05	2.09	2.78*	0.000	-	-	-
Empathy	*O*	0.002	-	-	-	0.002	-	-	-

$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 15: Result of statistical tests for LLaMA3.1 405B-low and LLaMA3.1 405B-high. Q columns indicate the Q-statistics from the Friedman test. Also, $\Delta_{i,j}$ columns show the score difference between i-th and j-th snapshots and corresponding post-hoc test results.