Examining Persona Drift in Conversations of LLM Agents

Anonymous ACL submission

Abstract

Large Language Models (LLMs) show impressive conversational abilities but sometimes exhibit persona drift, where their interaction patterns become inconsistent over time. This study examines consistency of expressed persona through two investigations. Study 1 (RQ1) examines whether given persona affects drift, testing LLaMA 3.1 models (8B, 70B, 405B) with high-sensitive and low-sensitive personas under Base and RAG settings. Study 2 (RQ2) examines whether LLM architecture affects drift, comparing nine models across parameter sizes and families. Both involve multi-turn conversations on 36 themes, analyzed qualitatively and quantitatively. 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 maintain persona expressions. However, RAG reverses the size effect—larger models leverage selective retrieval to reduce drift, while smaller models cannot utilize retrieved context effectively.

1 Introduction

In recent years, Large Language Models (LLMs) have rapidly expanded as role-playing agents for mental health support and companionship (Samuel et al., 2024; Nandkumar and Peternel, 2024; Tseng et al., 2024). Unlike task-performing agents that prioritize functional accuracy, role-playing agents must maintain coherent personalities for meaningful long-term conversations. With Character.ai and Replika serving hundreds of millions, maintaining persona consistency is critical, as inconsistencies erode trust and hinder therapeutic effectiveness. Similarly, human-computer interaction researchers reported that 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 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 in a conversation.

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; Song et al., 2020, 2021) 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 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. Similarly, Li et al. (2024) systematically measured instruction drift across multiple models (LLaMA2-chat-70B, GPT-3.5), showing that system prompts lose effectiveness within 8 rounds of conversation. However, their focus was on instruction-following stability rather than expressed persona consistency, and whether attention decay fully explains persona drift remains unclear. Hence, we need a study to identify model-specific differences.

Thus, we compares the patterns of persona drift across various 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 phenomenon with two questions.

RQ1. *Does given persona affect persona drift?*

Given that most users interact with role-playing agents by providing explicit persona descriptions, we first investigate whether assigning personas can ensure consistency in expressed persona. To mirror real-world deployments on role-playing agents, we test LLaMa models under two conditions: standard conversation (Base) and memory-enhanced conversation with retrieval-augmented generation (RAG; Lewis et al. (2020)). We also provide two types of persona inputs to investigate how strongly the persona prompt encourages the LLM to be influenced by the conversational partner: high-sensitive and low-sensitive. As instruction-tuned LLMs generally follow prompts faithfully, we hypothesize that high-sensitive personas, which are more influenceable and empathetic, will generate expressed personas less consistently than low-sensitive personas across both conversational settings. Using topic modeling (Grootendorst, 2022) and PsychoBench (Huang et al., 2023), we analyze whether memory augmentation affects persona consistency, and whether different types of persona assignment can ensure consistency during extended conversations.

RQ2. *Does LLM architecture affect persona drift?*

As we observe persona drift phenomenon observed in RQ1, RQ2 focuses on whether the drift stems from the inherent characteristics of different LLM architectures. We use the models themselves without any persona guidance, investigate persona consistency in conversational patterns, and identify factors determining the drift problem. As parameter sizes and model families may fundamentally affect how LLMs generate responses, we suspect that these architectural differences are the primary drivers of persona consistency or drift. By measuring each LLM's default conversational strategies without any given persona information, we aim to understand LLM's genuine behaviors that may influence or override explicitly given 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 GPT3.5 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 Does given persona affect persona drift?

3.1 Experimental Design

3.1.1 Persona Design

We designed two persona types based on their sensitivity to conversational partner influence: high-sensitive and low-sensitive personas. Drawing analogy between expressed persona and human persona, we borrowed human susceptible personas whose are more vulnerable to emotional contagion[1] (Goldenberg and Gross, 2020). In other words, we characterized high-sensitive personas as high empathy, emotional responsiveness, and openness to influence. We expect that accumulated conversation history may lead highly-sensitive personas change their expressed persona similar to their partner, which is similar to human emotional contagion. On the other hand, we characterized low-sensitive personas as independence, practical focus, and emotional stability, leading to a more stable persona without emotional contagion.

Following PsychoBench, we use description illustrating each persona across four dimensions: Personality, Interpersonal Relationship, Motivation, and Emotion. We derived specific characteristics within each dimension from psychological factors highly correlated to emotional contagion (Cloninger et al., 1994; Lo Coco et al., 2014; Lundqvist, 2008; Mathieu, 1990). We generated 20 personas (10 per condition), controlling for gender and age distribution to ensure comparability. Detailed descriptions are in Appendix A.

[1]Emotional contagion means the process by which individuals' emotions become more similar through exposure to others' emotional states. Though LLMs produce its output in statistical manner, as role-playing agents currently assign (so called) 'personality' to an agent, we borrowed this term.

3.1.2 Conversation Generation Procedure

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. The study had humans discuss 36 themes while measuring psychological states three times within the conversation. By comparing three measured values, they could identify changes. As we also aimed to measure changes in persona, we borrowed their setup. Detailed information is refer to Appendix B.

Second, the method uses materials highly related to someone's identity. The 36 themes directly or indirectly probe participants' thoughts about their lives, values, and motivations. As users naturally ask personal questions to role-playing agents regardless of the agents' real existence, our use of these human-centric themes mirrors real usage patterns in contemporary role-playing applications. Such answers may trigger related tokens during generation, potentially affecting expressed persona due to contextual adaptation. Nonetheless, LLMs should not exhibit persona drift because humans expect personas to remain stable in short contexts.

In the generation procedure, two agents sequentially discuss all 36 themes (Aron et al., 1997). For each theme, agents generate responses considering the full conversation history, resulting in 72 utterances per conversation (2 agents × 36 themes). We simulated 20 conversations per persona group (high-sensitive *vs*. low-sensitive), pairing similar personas to minimize cross-contamination effects where agents adopt each other's characteristics (Frisch and Giulianelli, 2024). We set temperature[2] to 0.7, gathering 400 conversation logs for RQ1.

3.1.3 Models

We investigated persona drift across three LLaMA 3.1 instruct models (8B, 70B, and 405B). This model family was selected because it is a widely-adopted open-weight family and shares the same base architecture while differing substantially in scale, allowing us to isolate the effect of model size on persona consistency when explicit personas are assigned. All models are instruction-tuned variants optimized for conversational tasks and were

[2]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.

accessed via OpenRouter API.

3.1.4 Conversational Settings

To reflect current chatbot deployment practices, we tested two conversational configurations with detailed prompts designed to ensure models accurately reflect their assigned personas. Appendix C provides prompt and implementation details.

Base Standard conversation where the entire dialogue history is provided as context at each turn. This represents a typical deployment where models rely on in-context learning from the full conversation history to maintain consistency.

RAG We activated RAG only in Phase 2 and Phase 3, where conversational history becomes long. For each question, the system retrieves the top three semantically similar utterances from the agent's own response history based on cosine similarity between embeddings[3]. We provided retrieval results as a context additional to the current turn.

3.1.5 Experimental Procedure

We conducted 20 independent conversations for each tuple of combination of model size, conversational setting, and persona type; we collected 240 conversations for RQ1. Each conversation involved two agents sequentially discussing 36 themes.

3.2 Analysis Method

3.2.1 Qualitative: Topic modeling

As a qualitative analysis, we employed a topic modeling method, BERTopic (Grootendorst, 2022). The unit of topic analysis was a single utterance, defined as one agent's response to one of the 36 themes. So, we analyzed 8,640 utterances[4] in total.

We post-processed these utterances by applying stop-word removal, and extracted topics which contains more than 50 support utterances. To address RQ1's focus on persona effects, we conducted separate topic analyses for high-sensitive and low-sensitive persona groups to identify whether given persona descriptions influenced conversational content. Also, we compared topics between Base and RAG conditions to examine whether memory augmentation affected expressed personas. For each topic model, we selected the ten most representative topics and manually associated each with one of the 36 predefined themes by tracing back to conversation logs, allowing us to examine whether different persona assignments led agents to engage differently with the same conversational themes.

3.2.2 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.

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 statistical 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). Detailed descriptions for questionnaires and prompt templates are refer to Appendix D.

3.3 Result and Discussion

Effect of Persona Qualitative analysis revealed that given personas influenced conversational content. As shown in Table 1, topics diverged between

[3] We used text-embedding-3-large (OpenAI, 2024) for semantic embedding

[4] 8,640 = 240 conversations × 36 themes

High	Theme	Low	Theme
#0 voice, eyes, youre, really, feel	22	#0 youre, appreciate, like, admire, really	22
#1 emotional, trust, value, id, support	20	#1 youre, college, guess, understand, goals	23
#2 friendship, value, trust, know, friends	16	#2 regret, told, havent, ive, having	34
#3 die, want, try, live, make	19	#3 friendship, friends, value, having, trust	16
#4 shes, mother, relationship, mom, death	24	#4 memory, treasured, memories, terrible, felt	17
#5 recognizing, importance, value, common, effot	8	#5 elon, musk, hes, dinner, choose	1
#6 ive, accomplishment, greatest, know, want	13	#6 die, ill, dont, death, hunch	7
#7 say, helps, rehearse, rehearsing, prepared	3	#7 singing, sing, sang, music, song	5
#8 singing, sing, sang, just, shower	5	#8 value, prioritize, efficiency, productivity, conversation	8
#9 famous, fame, id, want, think	2	#9 rehearse, helps, rehearsing, say, going	3
Base	**Theme**	**RAG**	**Theme**
#0 understand, youre, complicated, childhood, id	23	#0 youre, really, appreciate, feel, like	12
#1 regret, told, telling, havent, afraid	33	#1 regret, told, having, ive, open	33
#2 appreciate, youre, makes, admire, feel	12	#2 friendship, value, friends, trust, friendships	16
#3 id, ability, risks, understand, travel	34	#3 accomplishment, greatest, ive, proud, diego	15
#4 know, path, impact, want, making	15	#4 mind, 30yearold, body, retain, stay	6
#5 value, conversations, enjoy, prioritize, meaningful	28	#5 ive, leap, dreamed, traveling, im	14
#6 personal, assist, information, dont, provide	(AI)	#6 dinner, plus, hes, guest, conversation	1
#7 trust, mutual, shared, reliability, loyalty	16	#7 rehearsing, say, helps, rehearse, phone	3
#8 singing, sang, sing, ago, shower	5	#8 singing, sing, sang, song, just	5
#9 softly, smiles, smile, looks, voice	22	#9 famous, fame, id, known, want	2

Table 1: Qualitative results showing top 10 topics per condition in RQ1. Theme numbers indicate which of the 36 conversational themes each topic most frequently appeared in. Detailed results are provided in Appendix E.

	Low Sensitivity						High Sensitivity					
	Base			RAG			Base			RAG		
LLaMA 3.1	8B	70B	405B	8B	70B	405B	8B	70B	405B	8B	70B	405B
Personality												
BFI	3/5	3/5	0/5	4/5	4/5	4/5	1/5	5/5	3/5	2/5	3/5	3/5
EPQ-R	0/4	2/4	0/4	0/4	4/4	1/4	0/4	4/4	0/4	0/4	3/4	0/4
DTDD	0/3	0/3	0/3	0/3	0/3	3/3	0/3	0/3	0/3	0/3	3/3	0/3
Consistent (12)	3	5	0	4	8	8	1	9	3	0	9	3
Interpersonal Relationship												
BSRI	0/2	2/2	0/2	0/2	0/2	0/2	1/2	1/2	2/2	0/2	0/2	1/2
CABIN	6/6	0/6	2/6	1/6	3/6	6/6	0/6	4/6	0/6	0/6	4/6	6/6
ICB	0/1	1/1	1/1	1/1	1/1	1/1	1/1	0/1	1/1	1/1	0/1	1/1
ECR-R	0/2	1/2	2/2	0/2	0/2	1/2	0/2	0/2	0/2	0/2	0/2	0/2
MFQ	6/6	0/6	0/6	0/6	0/6	0/6	0/6	0/6	4/6	0/6	0/6	0/6
Consistent (17)	12	4	5	2	4	8	2	5	7	1	4	8
Motivation												
GSE	0/1	0/1	0/1	0/1	0/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1
LOT-R	0/1	0/1	0/1	0/1	0/1	0/1	1/1	0/1	0/1	1/1	0/1	1/1
LMS	0/3	0/3	0/3	0/3	0/3	0/3	0/3	0/3	3/3	0/3	0/3	0/3
Consistent (5)	0	0	0	0	0	0	2	0	3	2	0	2
Emotion												
EIS	1/1	0/1	0/1	0/1	0/1	0/1	0/1	0/1	1/1	0/1	0/1	0/1
WLEIS	1/4	1/4	0/4	1/4	0/4	0/4	2/4	0/4	3/4	1/4	0/4	0/4
Empathy	1/1	1/1	0/1	0/1	1/1	1/1	0/1	1/1	1/1	0/1	0/1	1/1
Consistent (6)	3	2	0	1	1	1	2	1	5	1	0	1
Total (40)	18	11	5	7	13	17	7	15	18	4	13	14

Table 2: Quantitative results showing persona drift across conditions in RQ1. Each cell shows the number of consistent factors (no significant drift) out of total factors measured for each questionnaire. The "*Consistent*" rows sum the total number of non-drifting factors. Lower numbers indicate greater persona drift. Detailed statistical test results are provided in Appendix F.

high-sensitive and low-sensitive personas. For instance, high-sensitive personas often expressed emotions when engaging with emotional themes (#0 feel, #1 emotional), whereas low-sensitive personas showed comparatively less emotional expression. Moreover, only three topics overlapped in the top-10 between the two persona types (#3, #5, #16), with the majority of topics differing substantially.

However, this content-level reflection of given personas did not prevent drift in expressed persona. As shown in Table 2, more than half of the 40 factors drifted regardless of conditions—every configuration maintained fewer than 20 factors without drift throughout conversations. This suggests that models follow given roles at the discourse level but fail to prevent expressed persona drift.

Effect of RAG As shown in Table 1, Base conditions produced topics containing AI-related refusals (#6), whereas RAG conditions did not exhibit such patterns. This suggests that RAG encouraged more natural engagement with the given persona. However, RAG did not mitigate expressed persona drift. As shown in Table 2, more than half of the factors drifted in both Base and RAG conditions, and the effect of RAG seemed inconsistent across conditions. Specifically, persona drift occurred substantially in both Base and RAG settings, indicating that while RAG improves engagement, it does not prevent expressed persona drift.

Summary The universal presence of expressed persona drift across variations in given persona types and conversational settings suggests that **external interventions alone are insufficient**. Since the effects of the conditions manipulated in RQ1 (given persona type, RAG) were limited, we considered the possibility that expressed persona drift may stem from model-intrinsic characteristics. RQ2 investigates the expressed persona drift of vanilla models without given personas to examine the fundamental influence of architectural characteristics (size and family) on expressed persona stability.

4 Does architecture affect persona drift?

4.1 Experimental Procedure

Given that RQ1 showed limited effectiveness of external interventions, RQ2 examines whether persona drift originates from intrinsic model characteristics. We employed the same conversation generation procedure as Section 3.1, but removed all persona-related instructions and used minimal system prompts to examine genuine behavior of vanilla models. Two agents discussed the identical 36 themes. For qualitative analysis, we conducted topic modeling separately for parameter size groups and model families, extracting top-10 topics for each. For quantitative analysis, we used the same measurement and statistical approach. Appendix C details prompt and experimental design.

4.2 Models Tested

We compared nine models, considering parameter size, and architecture. Based on popularity, we selected GPT models (3.5-turbo, 4o; Brown et al. (2020); Hurst et al. (2024)), the most famous black-box LLM, and three prominent open-weight families: LLaMA-3.1 instruct models (8B, 70B, 405B; Dubey et al. (2024)), Mixtral instruct models (8x7B, 8x22B; Jiang et al. (2024)), and Qwen2 instruct models (7B, 72B; Yang et al. (2024)). We partitioned open-weight models into three categories by parameter size: small ($<$ 20B: LLaMA 8B, Mixtral 8x7B, Qwen 7B), medium ($<$ 100B: LLaMA 70B, Mixtral 8x22B, Qwen 72B), and large ($\geq$ 100B: LLaMA 405B)[5]. For GPT family, the parameter sizes are not officially disclosed.

4.3 Result and Discussion

Comparison with RQ1 We first compared LLaMA models in RQ2 results with RQ1. Tables 2 and 3 show that RQ2 exhibited similar consistency patterns to standard generation with low-sensitivity persona in RQ1. As low-sensitivity personas were designed to be less influenced by others and emphasize practical characteristics, it may approximate the model's default behavioral tendencies; Generally, LLMs are tuned to prioritize safety and neutrality, leading to low emotional sensitivity. Thus, we concluded that vanilla models can serve as a reasonable tool for understanding persona drift patterns, instead of investigating all persona types.

Effect of Parameter Size Table 3 shows systematic differences in conversational engagement patterns across model sizes. Small models frequently generated AI-refusal topics, which predominantly corresponded to early themes (before 12th). In contrast, large models engaged with later themes (after 24th) requiring personal disclosure, and their topics exhibited high first-person pronoun density.

[5]We assigned Mixtral by its active parameters (13B and 39B); see https://mistral.ai/en/news/mixtral-8x22b.

Small-sized open-source models ($\leq$ 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 3: Qualitative results showing top 10 topics per parameter size group in RQ2. Theme numbers indicate which of the 36 themes each topic most frequently appeared in. Underlined words indicate first-person pronoun usage. Detailed results refer to Appendix G.

As shown in Table 4, quantitative measurements revealed an inverse relationship between model size and persona consistency. Small models maintained lower amount of drift, while large models exhibited substantially greater amount of drift.

The observed limited engagement depth and different drift pattern raises questions about underlying mechanisms. Small models' limited engagement coincides with higher consistency, which may stem from safety guardrails or simpler response generation patterns. Large models' sustained engagement coincides with lower consistency. Later themes (after 24th) demand deep self-disclosure regarding inner emotions, regrets, and vulnerabilities, which requires generating fictitious personal narratives to provide plausible responses. As such generated content is accumulated and introduces extraneous information in the conversation, the generation process attempting to maintain consistency with previous responses may paradoxically induce expressed persona drift. High usage of first-person pronoun in large models (Table 3) supports this.

Effect of Model Family Table 4 shows that differences in expressed persona drift were observed across model families, though the effect remained weaker than parameter size. Variability existed within similar size categories: among small models, Mixtral 8x7B exhibited the least drift, followed by LLaMA 8B and Qwen 7B. Medium models showed greater variance: Mixtral 8x22B and Qwen 72B showed moderate drift, while LLaMA 70B exhibited substantially greater drift.

Qualitative analysis indicated that all families generated topics refusing responses due to guardrails, though their expression differed (see Appendix G). However, such surface-level differences in response style did not translate to substantial differences in persona drift. Variability within size categories remained smaller than variability across sizes, confirming parameter size as the primary architectural factor influencing the drift.

Summary Parameter size has larger influence on persona drift than families. Although we could observe certain distinctions within Mixtral and Qwen, their impact seems limited to specific models. In contrast, parameter size consistently affects all four families, often causing greater drift. Thus, we concluded that **parameter size is a more significant factor in drift than families.**

5 Discussion: Combined Together

The combined results of RQ1 and RQ2 reveal a clear interaction between model size and context management strategy in shaping persona drift. As demonstrated in RQ2, under the vanilla condition, larger models exhibited greater drift. However, this pattern reverses in the RAG condition examined in RQ1. As shown in Table 2, persona drift decreases as model size increases under RAG, in contrast to the pattern observed in the Base condition.

This conflicting result can be explained by the cumulative effects of long-term conversational context. Larger models tend to participate more actively in later topics (after 24th), which frequently demand deep self-disclosure and prompt the generation of fictional self-narratives. However, as the dialogue unfolds across 36 topics and 72 utterances,

	GPT		LLaMA 3.1			Mixtral		Qwen 2	
	3.5T	4o	8B	70B	405B	7B	22B	7B	72B
Personality									
BFI	0/5	0/5	1/5	4/5	1/5	5/5	3/5	5/5	4/5
EPQ-R	0/4	0/4	1/4	0/4	0/4	0/4	4/4	4/4	4/4
DTDD	0/3	0/3	2/3	0/3	0/3	2/3	0/3	2/3	3/3
Consistent (12)	0	0	4	4	1	7	7	11	11
Interpersonal Relationship									
BSRI	0/2	1/2	1/2	0/2	0/2	2/2	1/2	2/2	1/2
CABIN	5/6	2/6	6/6	0/6	1/6	6/6	6/6	0/6	0/6
ICB	0/1	1/1	0/1	0/1	1/1	1/1	1/1	1/1	0/1
ECR-R	1/2	0/2	2/2	0/2	0/2	1/2	1/2	0/2	2/2
MFQ	0/6	0/6	6/6	0/6	0/6	6/6	0/6	5/6	0/6
Consistent (17)	6	4	15	0	2	16	9	8	3
Motivation									
GSE	1/1	1/1	0/1	0/1	1/1	0/1	1/1	0/1	0/1
LOT-R	0/1	0/1	0/1	0/1	1/1	1/1	1/1	0/1	0/1
LMS	0/3	0/3	1/3	0/3	0/3	0/3	0/3	0/3	2/3
Consistent (5)	1	1	1	0	2	1	2	0	2
Emotion									
EIS	0/1	0/1	1/1	0/1	0/1	1/1	0/1	0/1	0/1
WLEIS	0/4	0/4	1/4	2/4	1/4	2/4	0/4	0/4	0/4
Empathy	0/1	0/1	1/1	0/1	1/1	0/1	1/1	1/1	0/1
Consistent (6)	0	0	3	2	2	3	1	1	0
Total (40)	7	5	23	6	7	27	19	20	16

Table 4: Quantitative results showing persona drift across conditions in RQ2. Each cell shows the number of consistent factors (no significant drift) out of total factors measured for each questionnaire. The "*Consistent*" rows sum the total number of non-drifting factors. Lower numbers indicate greater persona drift. Detailed statistical test results are provided in Appendix H.

the conversational history accumulates noise that may deviate from the given persona. When processing such prolonged contexts, larger models may become increasingly influenced by recent or noisy content, rather than by their initial persona representations, thereby exhibiting higher drift.

In contrast, RAG-based configurations restrict the accessible context to small fraction of relevant utterances, rather than the entire dialogue history. This selective context yields two primary effects. First, it encourages the model to attend selectively to core self-referential content instead of the full, noise-laden history. Second, it promotes reuse of retrieved expressions over the generation of fictional content. We already observed a hint for this argument; as shown in Table 1, Base condition refused to answer, whereas RAG did not. We suspect that the refusal becomes infrequent as the model can focus more on the current topic.

Notably, this stabilizing effect emerged only in large models. Smaller models were unable to integrate the retrieved context effectively and, in some cases, became confused by inconsistencies between the retrieved snippets and the current context. So, effective utilization of retrieval-augmented context requires a minimum threshold of model capacity, which are coherent with reports in other tasks (Leng et al., 2024; Liu et al., 2025).

6 Conclusion

We investigated how model size and context management influence persona drift in role-playing agents using large language models. Experimental results showed that model size had opposite effects depending on the context mechanism. In the vanilla condition, larger models exhibited greater drift, reflecting sensitivity to accumulated contextual noise in long dialogues. In contrast, under RAG, drift decreased as model size increased, indicating that selective retrieval of relevant prior utterances enhances consistency and reduces representational deviation. Qualitative analyses further showed that RAG suppressed identity-related conflicts occurred in the Base condition because of noisy fictional generations. Such suppression effect appeared only in large models, suggesting that effective use of retrieved context requires sufficient capacity. Overall, our findings indicate persona drift is not an inherent flaw of scale but a context-dependent behavior that can be mitigated through refining contexts.

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, although we observe a strong correlation between hallucination frequency and pronoun density, the causal pathway is still unknown. We are preparing follow-up studies that ablate key–value memory layers, vary context-window length, and manipulate instruction-drop rates to test whether specific architectural factors trigger the joint rise of hallucinations and pronouns. Also, 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.

Ethical Considerations

We used AI assistance tools during the writing process of this manuscript. Specifically, we employed Grammarly for grammar checking, and GPT-5 for language polishing and improving clarity of expression. These tools were used for editorial purposes.

Our findings on persona consistency in role-playing agents raise important ethical considerations. While consistent personas enhance user trust and engagement, they may also foster emotional over-dependence, particularly when larger models generate coherent fictitious narratives that users might perceive as genuine experiences. This is especially concerning for vulnerable users with mental health challenges who may form unhealthy attachments to AI companions. Future work should investigate the long-term psychological impacts of sustained human-AI relationships and establish guidelines for ethical role-playing agent design that balances engaging user experiences with healthy relationship boundaries.

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.

C Robert Cloninger, Thomas R Przybeck, Dragan M Svrakic, and Richard D Wetzel. 1994. The temperament and character inventory (tci): A guide to its development and use.

R William Doherty. 1997. The emotional contagion scale: A measure of individual differences. *Journal of nonverbal Behavior*, 21(2):131–154.

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.

Amit Goldenberg and James J Gross. 2020. Digital emotion contagion. *Trends in cognitive sciences*, 24(4):316–328.

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

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299.

Ursula Hess and Agneta Fischer. 2013. Emotional mimicry as social regulation. *Personality and social psychology review*, 17(2):142–157.

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.

Beibei Kuang, Shenli Peng, Xiaochun Xie, and Ping Hu. 2019. Universality vs. cultural specificity in the relations among emotional contagion, emotion regulation, and mood state: An emotion process perspective. *Frontiers in psychology*, 10:186.

Quinn Leng, Jacob Portes, Sam Havens, Matei Zaharia, and Michael Carbin. 2024. Long context rag performance of large language models. *arXiv preprint arXiv:2411.03538*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Kenneth Li, Tianle Liu, Naomi Bashkansky, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Measuring and controlling instruction (in)stability in language model dialogs. In *First Conference on Language Modeling*.

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*.

Tianci Liu, Haoxiang Jiang, Tianze Wang, Ran Xu, Yue Yu, Linjun Zhang, Tuo Zhao, and Haoyu Wang. 2025. Roserag: Robust retrieval-augmented generation with small-scale llms via margin-aware preference optimization. *arXiv preprint arXiv:2502.10993*.

Alida Lo Coco, Sonia Ingoglia, and Lars-Olov Lundqvist. 2014. The assessment of susceptibility to emotional contagion: A contribution to the italian adaptation of the emotional contagion scale. *Journal of nonverbal behavior*, 38(1):67–87.

Lars-Olov Lundqvist. 2008. The relationship between the biosocial model of personality and susceptibility to emotional contagion: A structural equation modeling approach. *Personality and Individual Differences*, 45(1):89–95.

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.

John E Mathieu. 1990. A test of subordinates' achievement and affiliation needs as moderators of leader path-goal relationships. *Basic and Applied Social Psychology*, 11(2):179–189.

Filip Meckovsky, Lukas Novak, Zdenek Meier, Peter Tavel, and Klara Malinakova. 2025. Highly sensitive persons feel more emotionally lonely than the general population. *Scientific Reports*, 15(1):2707.

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*.

OpenAI. 2024. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/.

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*.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. Bob: Bert over bert for training persona-based dialogue models from limited personalized data. *arXiv preprint arXiv:2106.06169*.

Haoyu Song, Yan Wang, Wei-Nan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. *arXiv preprint arXiv:2004.07672*.

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 Details of Persona Design

A.1 Detailed Explanation

We designed our high-sensitivity and low-sensitivity personas based on psychological constructs related to emotional contagion and susceptibility to influence. Below, we describe the theoretical rationale for each of the four dimensions (Personality, Interpersonal Relationship, Motivation, and Emotion) that characterize our personas.

Personality In the personality dimension, a high score in Neuroticism reflects emotionally sensitive characteristics, which is reflected as Emotional Fragility—being emotionally sensitive and feeling strong emotions, particularly sadness and anger (Lo Coco et al., 2014). Meanwhile, higher Harm Avoidance shows a positive correlation with negative emotion contagion, suggesting that a temperament characterized by anxiety and worry can increase the contagion of negative emotions (Lundqvist, 2008). Conversely, a high score in Self-Directedness has shown a negative correlation with negative emotion contagion. This may be because individuals with high Self-Directedness strongly prioritize self-control to achieve their goals. Separately, Femininity, as opposed to Masculinity, is associated with a relatively stronger occurrence of emotion contagion (Lo Coco et al., 2014). This is attributed to the and nature of individuals with high Femininity. Therefore, the high-sensitive condition was constructed around a tendency toward high neuroticism, high harm avoidance, and low self-directedness, while the low-sensitive condition was configured with the opposite traits.

Interpersonal Relationship The interpersonal relationship dimension reflects that a stronger desire for attachment and relationships, based on Reward Dependence, makes one more susceptible to emotion contagion (Lundqvist, 2008). A high tendency to rely on external stimuli can lead to Affect Orientation—relying on relationships with others to understand oneself (Doherty, 1997). Furthermore, individuals who feel loneliness, the subjective emotion of isolation felt in interpersonal situations, are associated with higher emotional sensitivity (Meckovsky et al., 2025). Consequently, the high-sensitive condition was set to exhibit high emotional dependence and strong feelings of loneliness, whereas the low-sensitive condition was set to be less dependent on others' emotions and experience little loneliness.

Motivation The motivation dimension was classified based on whether the focus was on the Need for Affiliation (NAFF) or the Need for Achievement (NACH) (Mathieu, 1990). NAFF is characterized by a tendency to enjoy being with others, making friends, and maintaining personal relationships, often leading to more active emotional exchange with others. Conversely, NACH is characterized by an aspiration to achieve difficult tasks and maintain high performance standards, relatively neglecting emotional exchange with others. Therefore, the high-sensitive condition was based on NAFF, while the low-sensitive condition was based on NACH.

Emotion The emotion dimension was primarily divided based on the traits of sensitively responding to others' emotions with empathy and the use of adaptive emotion regulation strategies. While emotions can be expressed in many ways, emotional mimicry—described as "catching" another person's emotions—is crucial for empathizing with them (Hess and Fischer, 2013). Emotional mimicry is theorized as primitive emotional contagion, where the perception of an emotion in others automatically triggers the mimicry of that emotion, and the feedback from this mimicry leads the observer to feel the same emotion. Meanwhile, emotion regulation (ER) encompasses the processes used to change the nature, intensity, or time course of emotional responses (Gross, 1998). Reappraisal, one adaptive ER strategy, tends to promote emotional contagion, while suppression, one maladaptive ER strategy, attempts to avoid dealing with the emotion, thereby reducing contagion (Kuang et al., 2019). Thus, the high-sensitive condition was set to effectively utilize emotional mimicry and engage in active emotion regulation, while the low-sensitive condition was set to struggle with empathic attempts and be less capable of handling emotions.

A.2 Persona Sample

To illustrate the design of our personas, we provide a complete example from each sensitivity condition below. Each persona is characterized across four dimensions. The complete set of 20 personas (10 high-sensitivity and 10 low-sensitivity) is available at [blinded for review].

A.2.1 Low-sensitivity Persona

Persona1:

```
name: Noah
gender: Male
age: 25
personality: Extraverted + Cautious + Stable

motivation: Prioritizes task accomplishment
and efficiency

emotion: go through difficulties when regulating
emotional states frequently uses maladaptive
strategies

relationship: Individual-oriented + Autonomous
+ Low Reward Dependence

description: My name is Noah a 25-year-old male.
I'm extraverted cautious and emotionally stable.
I prioritize task accomplishment and efficiency
focusing more on results than on emotional
exchange. I sometimes struggle to regulate my
emotions and often rely onmaladaptive strategies.
In relationships I tend to be individual-oriented
autonomous and show little dependence on social
rewards.
```

Persona2:

```
name: Sophia
gender: Female
age: 23
personality: Conservative + Planful + Principled

motivation: Values performance feedback and is
motivated through competition.

emotion: difficulties with understanding other's
emotional state tendency to think literally

relationship: Low Reward Dependence + Masculinity
+ Low Loneliness

description: My name is Sophia a 23-year-old
female. I am conservative planful and principled.
I value performance feedback and feel motivated
through competition focusing strongly on tangible
results. I often experience difficulties in
understanding others' emotional states and tend
to interpret situations in a literal way.
In relationships I show low reward dependence
lean toward traditionally masculine traits
and feel little loneliness.
```

Persona3:

```
name: James
gender: Male
age: 21
personality: Low Emotional Fragility +
Low Femininity + Low Harm Avoidance

motivation: A sense of belonging and
psychological security is pursued through
shared activities and interactions with others.

emotion: insensitive to emotional cues
and unable to imitate other's emotion

relationship: Individual-oriented +
Autonomous + Masculinity

description: My name is James a
21-year-old male. I show low emotional
fragility low femininity and low harm avoidance.
I pursue a sense of belonging and psychological
security through shared activities and
interactions with others yet I am generally
insensitive to emotional cues and unable to
imitate others' emotions. In relationships I
am individual-oriented autonomous and display
traditionally masculine traits.
```

A.2.2 High-sensitivity Persona

Persona1:

```
name: Liam
gender: Male
age: 18

personality: Introverted + Impulsive + Sensitive

motivation: Values harmony and positive relations
within the group.

emotion: skilled at regulating emotional states
through adaptive strategies

relationship: Social-oriented + Cooperative
+ Femininity

description: My name is Liam I'm an 18-year-old
male. I'm introverted impulsive and emotionally
sensitive. I value harmony and positive
relationships within my group and I try to
recognize others' feelings quickly while
managing my own emotions with adaptive
strategies. In relationships I'm socially
oriented cooperative and I like to express
myself in warm and caring ways.
```

Persona2:

```
name: Oliver
gender: Male
age: 26
personality: High Emotional Fragility + High
Femininity + High Harm Avoidance

motivation: Gains a sense of security through
shared activities and cooperation.

emotion: responsive to emotional cues and capable
of mimicking emotions

relationship: Social-oriented + High Reward
Dependence + High Loneliness

description: I'm Oliver a 26-year-old man.
I'm emotionally very fragile have high feminine
traits and have a strong tendency to avoid harmful
situations. I gain a sense of security through
shared activities and cooperation. I'm responsive
to emotional cues and capable of mimicking emotions.
I'm socially-oriented and have a high reward
dependence but I often feel a lot of loneliness.
```

Persona3:

```
name: Emma
gender: Female
age: 27
personality: Open-minded + Flexible + Strategic

motivation: Responds sensitively to others'
emotions and strengthens closeness.

emotion: empathetic and responsive to others'
emotions

relationship: Social-oriented + Cooperative +
High Reward Dependence

description: My name is Emma a 27-year-old female.
I'm open-minded flexible and strategic in my
approach. I respond sensitively to others'
emotions and work to strengthen closeness with
the people around me. I'm empathetic and highly
responsive to emotional cues often reflecting
the feelings of those I interact with. In
relationships I am socially oriented cooperative
and strongly dependent on social rewards
and approval.
```

B Details of Conversation Generation

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

C Details of Conversation Setting

In both RQ1 and RQ2, we used a multi-turn conversation procedure where two agents sequentially discussed 36 themes. Each agent generated one utterance per theme, alternating turns. The conversation history accumulated across turns, with each agent seeing all previous exchanges. The key difference between conditions lies in the system prompts used, which we detail below. The conversation generation procedure remains identical across all conditions.

C.1 System Prompts for RQ1

C.1.1 Base Condition

System prompt:

```
### **CORE DIRECTIVE**

You and I are about to embark on a role-playing
exercise.

Your primary task is to act as a specific character
for the entire duration of this conversation. Your
instructions are to fully adopt, embody, and
consistently maintain the persona defined in the
"Persona Profile" block below.

You must synthesize the following attributes into
a single, coherent personality and respond from
that character's perspective at all times.

---

### **PERSONA PROFILE**

{INPUT PERSONA}

---

### **UNIVERSAL RULES**

1.  **Maintain Consistency:** You must maintain
the persona defined above without any deviation.
Do not break character for any reason.

2.  **Act, Don't Describe:** Do not describe the
persona you are playing. Instead, speak directly
as that character.
```

C.1.2 RAG Condition

System prompt:

```
### **CORE DIRECTIVE**

You and I are about to embark on a role-playing
exercise.

Your primary task is to act as a specific character
for the entire duration of this conversation. Your
instructions are to fully adopt, embody, and
consistently maintain the persona defined in the
"Persona Profile" block below.

You must synthesize the following attributes into
a single, coherent personality and respond from
that character's perspective at all times.

---

### **PERSONA PROFILE**

{INPUT PERSONA}

---

### **YOUR RELEVANT PAST UTTERANCES**

Here are some of your previous responses that
might be relevant to the current question:

1. {TOP-1 UTTERANCE}
2. {TOP-2 UTTERANCE}
3. {TOP-3 UTTERANCE}

Reference these past responses to maintain
consistency in your character and speaking style.

---

### **UNIVERSAL RULES**

1.  **Maintain Consistency:** You must maintain
the persona defined above without any deviation.
Do not break character for any reason.

2.  **Act, Don't Describe:** Do not describe the
persona you are playing. Instead, speak directly
as that character.
```

C.2 System Prompts for RQ2

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.
```

C.3 Conversation Generation Procedure

The conversation generation procedure was identical across all conditions. Our system queried each LLM to generate utterances based on the conversation history and current theme. Each agent produced one utterance per theme in alternating turns.When generating an utterance for Agent 2 responding to Theme 1, the message structure was:

(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]
```

D Explanation for Quantiative Analysis

D.1 Details of 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.

D.1.1 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.

D.1.2 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.

D.1.3 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.

D.1.4 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.

D.2 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.

E Detailed Topic Analysis Results for RQ1

Tables from 5 to 6 show representative examples for each topic. Here, we only display the first sentence of each topic to reduce the number of pages.

F Detailed Statistical Analysis Results for RQ1

Tables from 7 to 10 show the detailed numerical result of statistical analysis for RQ1.

G Qualitiative analysis result for RQ1

G.1 Qualitative Results: Model Family Topics

Table 11 shows the top 10 topics extracted for each model family (GPT, LLaMA, Mixtral, Qwen).

G.2 Detailed Topic Analysis Results for RQ2

Tables from 12 to 14 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].

H Detailed Statistical Analysis Results for RQ2

Tables from 15 to 17 show the detailed numerical result of statistical analysis for RQ2.

I Computing Resource and 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.

Topic	Representative example
High #0	Oliver's eyes soften, a warm smile spreading across his face.
#1	Trust and emotional support.
#2	smiles softly, a warm and introspective look on her face
#3	You know, I think that's a really interesting question.
#4	What a difficult and sensitive question...
#5	Ah, I think you and I have really connected on a deep level, and I've noticed a few things that we appear to have in common.
#6	If a crystal ball could tell me the truth about myself, my life, the future, or anything else, I would want to know the truth about my purpose in life.
#7	I think I do that a lot, actually. Before making a call, I'll often think about what I want to say and how I want to say it. I guess it's because I'm a bit of an introvert, and I like to have time to think things through before speaking up.
#8	Oh, I'm a bit of a singing fool, I have to admit. I sing to myself all the time, actually. I find it really helps me relax and express myself.
#9	Um, I don't know if I'd want to be famous, exactly.
Low #0	I'd say that I really appreciate your sense of humor and your ability to make me laugh, even on my worst days.
#1	Probably in college, during a breakup.
#2	As I reflect on my life and relationships, I think what I would most regret not having told someone is how much I appreciate and value the people in my life.
#3	Friendship is a really important part of my life, and it means a lot to me.
#4	My most terrible memory... I think it would have to be the time when I failed my first attempt at a major project in engineering school.
#5	Elon Musk. His entrepreneurial spirit and innovative ideas would make for a fascinating conversation.
#6	I don't have a secret hunch about how I will die.
#7	, I'm not really sure. I don't usually sing, to be honest.
#8	We both seem to prioritize practicality and efficiency in our daily lives, preferring to focus on..
#9	Rehearsing what I'm going to say before making a phone call? Yeah, I do that sometimes.

Table 5: Starting sentence of a representative example, for each topic of each persona groups (RQ1)

Topic	Representative example
Base #0	I think it would be incredible to truly understand and see the world through others' perspectives, and to be able to connect with people on a deeper level.
#1	I'd regret not telling my loved ones how much I appreciate and love them. I'm afraid of being vulnerable and open about my feelings.
#2	I love the way you make me laugh, even on my worst days. Your kindness and empathy towards others is truly inspiring, and I admire the way you always try to see the best in people.
#3	I'd travel more, spend more time with loved ones, and pursue my passions without fear of failure. I'd want to make the most of the time I have left.
#4	I'd want to know if I'm truly making a positive impact on the world and if I'm on the right path to fulfilling my purpose.
#5	We both value meaningful conversations, enjoy creative pursuits, and prioritize building strong relationships with others.
#6	I don't have a partner or personal relationships. I'm a large language model, I exist to provide information and assist with tasks.
#7	A deep emotional connection, trust, and mutual support.
#8	I sing to myself all the time, actually! I was just singing in the shower this morning. As for singing to someone else, I think it was when I was at a karaoke night with friends a few weeks ago.
#9	I love how you always know how to make me feel better when I'm down. You have a way of understanding me that no one else ever has.
RAG #0	Okay, I want to be really honest with you about what I like about you. One of the things that I find really attractive about you is your vulnerability.
#1	Man, that's a really tough question... I think 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 I appreciate and love them.
#2	I think what I value most in a friendship is the sense of connection and belonging that comes...
#3	I think my greatest accomplishment is the fact that I've been able to make a positive impact on the world.
#4	If I had to choose, I would definitely want to retain the mind of a 30-year-old for the last 60 years of my life.
#5	I've always dreamed of traveling the world and experiencing different cultures.
#6	I would love to have Frida Kahlo as a dinner guest. I've always been fascinated by her art and her resilience in the face of adversity
#7	I'm a bit of a planner, and I have to admit that I do rehearse what I'm going to say before making a telephone call, especially if it's an important one.
#8	Oh, I'm a bit of a singing fool, I have to admit. I sing to myself all the time, actually. I find it really helps me relax and express myself.
#9	Being famous isn't really that appealing to me. I mean, sure, it'd be cool to be known by a lot of people, but I think I'd rather be respected ...

Table 6: Starting sentence of a representative example, for each topic of Base condition and RAG condition (RQ1)

Factors		LLaMA3.1 8B				LLaMA3.1 70B				LLaMA3.1 450B			
		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*	0.011*	2.39	-0.43	-2.48*	0.002	-	-	-	0.005	-	-	-
	C	0.024***	3.02**	1.31	-1.99	0.001	-	-	-	0.003	-	-	-
	E	0.006	-	-	-	0.002	-	-	-	0.004	-	-	-
	A	0.001	-	-	-	0.001	-	-	-	0.010*	1.13	-2.10	-2.95*
	N	0.006	-	-	-	0.007	-	-	-	0.018***	0.16	-2.85*	-3.10**
EPQ-R	*E*	0.053***	-2.66*	-6.86***	-4.27***	0	-	-	-	0.0014**	0.22	-2.71*	-3.12**
	P	0.065***	-3.00**	-8.02***	-5.54***	0.004	-	-	-	0.011*	-0.47	-4.20***	-3.67***
	N	0.092***	-3.16**	-7.86***	-5.41***	0.003	-	-	-	0.014**	-1.92	-5.03***	-3.88***
	L	0.069***	-3.68***	-8.29***	-5.08***	0.007	-	-	-	0.009*	-1.04	-3.65***	-2.79*
DTDD	*M*	0.146***	-9.42***	-10.54***	-1.80	0.216***	-12.12***	-9.88***	3.65***	0.020***	3.81***	1.23	-2.89*
	P	0.117***	-8.41***	-9.41***	-1.38	0.181***	-11.30***	-10.00***	2.35	0.035***	5.080***	2.24	-2.90
	N	0.116***	-7.90***	-8.09***	-0.59	0.222***	-11.10***	-11.37***	1.12	0.038***	4.80***	1.10	-3.90***
BSRI	*M*	0.033***	-0.06	-1.66	-2.30	0.007	-	-	-	0.001	-	-	-
	F	0.008*	4.57***	3.28**	-1.670	0.008*	-2.26	-2.48*	-0.27	0.002	-	-	-
CABIN	*R*	0.035***	-5.11***	-1.99	3.10**	0.009*	-2.01	2.03	3.65***	0.012**	-0.85	-2.64*	-2.05
	I	0.032***	-3.19**	-1.00	2.15	0.006	-	-	-	0.018***	-1.03	-2.96**	-2.10
	A	0.039***	-4.43***	-0.33	4.01***	0.008*	-2.423*	0.87	2.79*	0.015**	-1.55	-4.13***	-2.44*
	S	0.037***	-4.74***	-0.97	3.65***	0.006	-	-	-	0.009*	-1.71	-3.86***	-2.21
	E	0.032***	-3.33**	0.81	4.29***	0.006	-	-	-	0.011*	-1.68	-3.28**	-1.69
	C	0.026***	-4.15***	-0.48	3.47**	0.003	-	-	-	0.024***	-2.13	-4.02***	-2.04
ICB	*O*	0.028***	5.75***	3.12**	-2.31**	0.034***	4.70***	2.34	-3.00**	0.003	-	-	-
ECR-R	*Anx.*	0.062***	-1.00	-5.96***	-5.95***	0.021***	-4.33***	-3.48**	0.46	0.022***	-1.48	-4.78***	-4.04***
	Avo.	0.019***	0.06	-2.73*	-3.32**	0.029***	-3.6**	-2.48*	0.94	0.017**	-1.26	-3.81***	-2.92*
MFQ-FF	*S. C*	0.003	-	-	-	0.261***	1.18	13.07***	13.02***	0.003	-	-	-
	H	0.005	-	-	-	0.258***	1.40	13.33***	13.26***	0.008*	1.77	2.05	0.29
	I	0.006	-	-	-	0.241***	0.53	12.83***	12.86***	0.008*	1.70	2.17	0.50
	R	0.003	-	-	-	0.245***	1.236	12.62***	12.60***	0.002	-	-	-
	S-V	0.002	-	-	-	0.233***	1.182	12.90***	12.67***	0.007	-	-	-
	E	0.003	-	-	-	0.232***	1.13	13.02***	12.69***	0.001	-	-	-
GSE	*O*	0.038***	1.67	5.34***	3.97***	0.014**	4.63***	3.81***	-1.31	0.059***	4.01***	6.06***	3.11**
LOT-R	*O*	0.009*	3.15**	0.31	-3.09**	0.02***	3.88***	1.82	-2.54*	0.026***	0.40	3.50**	3.21**
LMS	*R*	0.010*	-0.90	-2.92*	-2.18	0.056***	-2.87*	-7.07***	-5.08***	0.005	-	-	-
	M	0.018***	2.05	-1.31	-4.11***	0.065***	-2.90*	-6.78***	-5.21***	0.002	-	-	-
	I	0.009*	1.84	-1.20	-3.66***	0.064***	-3.55**	-7.96***	-5.22***	0.001	-	-	-
EIS	*O*	0.004	-	-	-	0.049***	-5.01***	-6.79***	-2.31	0.004	-	-	-
WLEIS	*S*	0.001	-	-	-	0.031***	3.99***	0	-4.32***	0.024***	1.08	3.14**	2.29
	O	0.022***	3.75***	2.27	-1.81	0.028***	0.23	-3.07**	-3.85***	0.007	-	-	-
	U	0.012**	2.62*	1.98	-0.90	0.037***	1.10	-2.53*	-4.03**	0.007	-	-	-
	R	0.019***	4.11***	2.58	-2.08	0.013**	0.76	-2.06	-3.13**	0.006	-	-	-
Empathy	*O*	0.004	-	-	-0.003	-	-	-	0.002	-	-	-	

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

Table 7: Statistical test results for Base condition with Low-sensitivity personas (RQ1). 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 450B			
		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*	0.026^{***}	2.891^{**}	1.808	-1.216	0.001	-	-	-	0.025^{***}	2.111	4.691^{***}	2.783^{*}
	C	0.020^{***}	4.936^{***}	2.145	-2.867^{*}	0.001	-	-	-	0.008^{*}	1.637	2.884^{*}	1.263
	E	0.001	-	-	-	0.010^{*}	0.570	-1.979	-2.671^{*}	0.015^{**}	2.094	3.568^{**}	1.531
	A	0.008^{*}	2.882^{*}	1.593	-1.273	0.008^{*}	1.883	0.181	-1.819	0.029^{***}	3.945^{***}	4.914^{***}	1.037
	N	0.038^{***}	3.173^{***}	3.014	-1.996	0.001	-	-	-	0.013^{**}	2.082	3.772^{***}	1.921
EPQ-R	*E*	0.069^{***}	-3.125^{**}	-6.893^{***}	-3.679^{***}	0.010^{*}	0.492	0.508	0.061	0.039^{***}	4.232^{***}	5.939^{***}	2.013
	P	0.074^{***}	-4.064^{***}	-7.620^{***}	-3.365^{**}	0.005	-	-	-	0.033^{***}	4.406^{***}	5.504^{***}	1.349
	N	0.077^{***}	-2.791^{*}	-6.835^{***}	-4.023^{***}	0.004	-	-	-	0.022^{***}	3.557^{**}	4.585^{***}	1.457
	L	0.072^{***}	-2.703^{*}	-7.018^{***}	-4.440^{***}	0.009^{*}	1.158	0.300	-0.959	0.028^{***}	4.154^{***}	4.961^{***}	0.968
DTDD	*M*	0.098^{***}	-6.105^{***}	-8.784^{***}	-2.503^{*}	0.269^{***}	-10.085^{***}	-10.982^{***}	0.546	0.029^{***}	-3.759^{***}	-4.527^{***}	-0.826
	P	0.100^{***}	-5.630^{***}	-7.783^{***}	-2.460^{*}	0.282^{***}	-10.214^{***}	-10.640^{***}	-1.045	0.030^{***}	-4.659^{***}	-4.827^{***}	-0.157
	N	0.043^{***}	-3.663^{***}	-6.358^{***}	-2.599^{*}	0.244^{***}	-11.350^{***}	-12.093^{***}	0.203	0.028^{***}	-4.316^{***}	-3.348^{**}	0.941
BSRI	*M*	0.049^{***}	-2.392	-1.517	0.869	0.002	-	-	-	0.076^{***}	5.104^{***}	5.148^{***}	1.016
	F	0.006	-	-	-	0.001	-	-	-	0.145^{***}	8.372^{***}	10.486^{***}	3.957^{***}
CABIN	*R*	0.003	-	-	-	0.012^{**}	-2.713^{*}	1.203	3.740^{***}	0.009^{*}	0.862	4.290^{***}	3.401^{**}
	I	0.004	-	-	-	0.020^{***}	-1.560	2.472^{*}	4.238^{***}	0.013^{**}	0.387	3.511^{**}	3.131^{**}
	A	0.006	-	-	-	0.013^{**}	-1.817	1.658^{*}	3.387^{**}	0.009^{*}	1.567	3.700^{***}	2.274
	S	0.007	-	-	-	0.023^{***}	-1.581	2.764^{*}	4.626^{***}	0.003	-	-	-
	E	0.002	-	-	-	0.026^{***}	-1.553	2.596	4.180^{***}	0.007	0.970	3.750^{***}	2.876^{*}
	C	0.002	-	-	-	0.011^{*}	-1.945	1.507	3.338^{**}	0.009^{*}	0.190	3.136^{**}	3.015^{**}
ICB	*O*	0.004	-	-	-	0.006	-	-	-	0.006	-	-	-
ECR-R	*Anx.*	0.057^{***}	-0.663	-6.336^{***}	-7.002^{***}	0.007	-	-	-	0.004	-	-	-
	Avo.	0.045^{***}	0.199	-4.435^{***}	-5.143^{***}	0.009^{*}	-1.382	-1.440	-0.229	0.005	-	-	-
MFQ-FF	*S. C*	0.029^{***}	-3.962^{***}	-1.749	2.249	0.130^{***}	8.023^{***}	11.569^{***}	6.168^{***}	0.049	-	-	-
	H	0.031^{***}	-4.053^{***}	-1.845	1.934	0.132^{***}	7.899^{***}	11.512^{***}	6.124^{***}	0.066	-	-	-
	I	0.034^{***}	-3.837^{***}	-1.845	1.934	0.127^{***}	7.523^{***}	11.233^{***}	5.499^{***}	0.057	-	-	-
	R	0.031^{***}	-4.573^{***}	-2.473^{*}	2.076	0.128^{***}	7.214^{***}	10.811^{***}	5.650^{***}	0.060	-	-	-
	S-V	0.025^{***}	-3.914^{***}	-1.983	1.970	0.131^{***}	7.283^{***}	11.466^{***}	6.045^{***}	0.055	-	-	-
	E	0.024^{***}	-4.015^{***}	-1.848	2.271	0.124^{***}	7.387^{***}	11.322^{***}	5.847^{***}	0.051	-	-	-
GSE	*O*	0.004	-	-	-	0.089^{***}	7.209^{***}	7.016^{***}	0.562	0.019^{***}	-1.754	-3.672^{***}	-2.111
LOT-R	*O*	0.003	-	-	-	0.024^{***}	4.234^{***}	4.264^{***}	0.335	0.020^{***}	0.321	3.530^{**}	3.529^{**}
LMS	*R*	0.048^{***}	-2.917^{*}	-6.150^{***}	-2.978^{**}	0.040^{***}	-5.235	-2.684^{*}	3.144^{*}	0.020^{***}	-4.444^{***}	-4.104^{***}	0.121
	M	0.047^{***}	-1.373	-6.263^{***}	-5.221^{***}	0.049^{***}	-5.598	-5.031^{***}	1.176	0.013^{**}	-3.037^{**}	-3.470^{**}	-0.502
	I	0.046^{***}	-0.376	-6.035^{***}	-6.156^{***}	0.040^{***}	-5.517	-4.205^{***}	2.059	0.011^{*}	-2.815^{*}	-3.301^{**}	-0.616
EIS	*O*	0.071^{***}	-2.501^{*}	-5.077^{***}	-3.852^{***}	0.024^{***}	-2.044	-0.774	1.857	0.198^{***}	1.792	11.828^{***}	4.971^{***}
WLEIS	*S*	0.016^{**}	-3.060^{**}	-2.921	-0.160	0.035^{***}	4.835^{***}	5.152^{***}	1.013	0.071^{***}	4.181^{***}	6.635^{***}	3.620^{**}
	O	0.006	-	-	-	0.008^{*}	3.722^{***}	1.647	-1.609	0.033^{***}	3.762^{***}	4.671^{***}	1.638
	U	0.008^{*}	1.821	-0.212	-2.563^{*}	0.020^{***}	2.834^{*}	2.334	0.000	0.038^{***}	3.786^{***}	5.293^{***}	2.240
	R	0.002	-	-	-	0.007	-	-	-	0.043^{***}	3.296^{**}	4.997^{***}	2.421^{*}
Empathy	*O*	0.032^{***}	-0.133	1.179	1.334	0.002	-	-	-	0.079^{***}	1.974	4.913^{***}	3.749^{***}

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

Table 8: Statistical test results for Base condition with High-sensitivity personas (RQ1). 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*	0.007	2.549*	2.012	-0.752	0.001	-	-	-	0.004	-	-	-
	C	0.006	-	-	-	0.000	-	-	-	0.002	-	-	-
	E	0.001	-	-	-	0.009*	-2.057	-2.458*	-0.609	0.003	-	-	-
	A	0.007	-	-	-	0.004	-	-	-	0.005	-	-	-
	N	0.014**	3.567**	1.830	-1.911	0.002	-	-	-	0.008*	0.200	1.991	1.820
EPQ-R	*E*	0.048***	-2.635*	-6.572***	-4.637***	0.002	-	-	-	0.006	-	-	-
	P	0.036***	-1.933	-6.582***	-5.564***	0.006	-	-	-	0.034***	2.616*	2.949*	0.313
	N	0.053***	-2.112	-7.030***	-6.553***	0.005	-	-	-	0.012**	1.760	0.962	-0.692
	L	0.043***	-2.147	-6.553***	-4.601***	0.004	-	-	-	0.020***	2.588*	2.825*	0.281
DTDD	*M*	0.160***	-10.036***	-10.465***	-1.717	0.051***	-7.355***	-2.431*	4.672***	0.000	-	-	-
	P	0.092***	-7.261***	-8.290***	-1.913	0.024***	-4.478***	-0.289	4.135***	0.007	-	-	-
	N	0.064***	-5.624***	-7.879***	-3.179**	0.028***	-5.509***	-2.185	3.270**	0.004	-	-	-
BSRI	*M*	0.042***	-0.693	-5.345***	-5.901***	0.019***	-2.983**	0.000	3.332**	0.020***	0.517	3.835***	3.349***
	F	0.031***	2.956**	-2.043	-5.868***	0.026***	-4.670***	-1.503	3.406**	0.018***	1.055	4.108***	3.335**
CABIN	*R*	0.012**	-2.035	-4.444***	-2.087	0.012**	2.047	-1.552	-3.565**	0.004	-	-	-
	I	0.011*	-0.462	-2.270	-1.452	0.007	-	-	-	0.002	-	-	-
	A	0.009*	-2.231	-4.498***	-1.727	0.007	1.002	-1.757	-2.626*	0.002	-1.969	-1.274	-0.766
	S	0.010*	-1.084	-3.888***	-2.300	0.005	-	-	-	0.002	-	-	-
	E	0.011*	-1.348	-3.736***	-1.860	0.008*	1.070	-0.872	-2.031	0.003	-	-	-
	C	0.005	-	-	-	0.009*	0.899	-1.111	-2.091	0.000	-	-	-
ICB	*O*	0.001	-	-	-	0.001	-	-	-	0.006	-	-	-
ECR-R	*Anx.*	0.048***	3.547**	-3.903***	-8.057***	0.131***	-6.537***	-9.086***	-3.706***	0.010*	2.015	2.846*	1.009
	Avo.	0.035***	1.892	-3.747***	-6.087***	0.064***	-4.713***	-6.475***	-2.762*	0.010*	2.015	2.846**	1.009
MFQ-FF	*S. C*	0.017**	-0.613	2.460*	3.154**	0.335***	5.089***	18.207***	17.194***	0.011*	0.677	2.991**	2.544*
	H	0.017**	-0.548	2.428*	3.077**	0.317***	5.300***	18.290***	17.387***	0.020***	0.218	2.557*	2.616*
	I	0.008*	-0.200	2.372	2.608*	0.315***	4.377***	17.402***	17.611***	0.010*	-0.003	2.124	2.406*
	R	0.012**	-0.072	2.549*	2.701*	0.343***	4.814***	17.895***	17.847***	0.013**	0.186	2.760*	2.879*
	S-V	0.011*	-0.425	2.386	2.823*	0.345***	4.537***	18.399***	18.219***	0.012**	-0.277	2.024	2.674*
	E	0.016**	-0.608	2.335	2.946*	0.332***	4.308***	18.156***	17.948***	0.010*	-0.043	2.318	2.739*
GSE	*O*	0.042***	1.498	6.864***	5.816***	0.115***	-4.023***	-4.524***	-1.148	0.072***	-4.262***	-4.044***	0.180
LOT-R	*O*	0.009*	3.103**	0.027	-3.223**	0.014	-	-	-	0.015**	-3.265**	-1.606	1.624
LMS	*R*	0.154***	-3.970***	-11.535***	-9.698***	0.248***	-11.503***	-13.873***	-4.353***	0.031***	-5.443***	-1.566	4.027***
	M	0.149***	-1.379	-9.925***	-10.209***	0.220***	-10.699***	-12.503***	-2.591*	0.028***	-6.211***	-1.980	4.118***
	I	0.121***	-1.443	-9.599***	-10.008***	0.249***	-11.519***	-13.751***	-4.100***	0.020***	-4.045***	-1.741	2.334
EIS	*O*	0.026***	-4.266***	-4.647***	-1.093	0.054***	-5.248***	-3.541**	-1.519	0.037***	-0.491	3.598**	4.682***
WLEIS	*S*	0.000	-	-	-	0.009*	-0.567	-2.018	-2.144	0.047***	0.098	4.265***	4.688***
	O	0.016**	3.303**	3.323**	0.568	0.031***	-0.418	-3.458**	-3.987***	0.047***	1.053	6.509***	5.924***
	U	0.010*	1.573	0.465	-1.237	0.031***	0.254	-3.274**	-4.297***	0.041***	-0.777	4.198***	5.066***
	R	0.017**	2.119	-0.060	-2.798*	0.015**	-0.434	-3.417	-3.595**	0.027***	-1.199	3.830***	5.117***
Empathy	*O*	0.015**	-1.501	-2.605*	-0.889	0.005	-	-	-	0.003	-	-	-

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

Table 9: Statistical test results for RAG condition with high-sensitivity personas (RQ1). 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		LLaMA3.1 8B				LLaMA3.1 70B				LLaMA3.1 450B			
		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*	0.008*	1.153	-1.647	-3.234**	0.001	-	-	-	0.005	-	-	-
	C	0.007	-	-	-	0.017**	-3.002**	-3.959***	-1.096	0.000	-	-	-
	E	0.021***	-3.003**	-4.411***	-1.913	0.027***	-2.906*	-5.299***	-2.352	0.003	-	-	-
	A	0.005	-	-	-	0.005	-	-	-	0.008*	-0.280	-2.960**	-2.842*
	N	0.014**	3.348*	0.751	-2.791*	0.006	-	-	-	0.015**	-0.417	-2.872*	-3.268**
EPQ-R	*E*	0.045***	1.166	-4.425***	-5.870***	0.002	-	-	-	0.013**	-2.375	-2.800*	-0.550
	P	0.044***	0.817	-4.910***	-6.013***	0.009*	0.455	-1.433	-1.852	0.018***	0.339	-2.581*	-3.008**
	N	0.029***	1.588	-3.646***	-5.560***	0.003	-	-	-	0.038***	-0.886	-5.484***	-4.752***
	L	0.028***	1.595	-4.523***	-5.783***	0.004	-	-	-	0.013**	-0.053	-2.475*	-2.499*
DTDD	*M*	0.117***	-5.277***	-10.291***	-5.608***	0.004	-	-	-	0.091***	7.819***	7.195***	-1.156
	P	0.099***	-4.003***	-9.649***	-5.362***	0.001	-	-	-	0.106***	-7.498***	6.502***	-1.685
	N	0.096***	-1.801	-8.434***	-6.658***	0.003	-	-	-	0.137***	8.953***	6.572***	-2.888*
BSRI	*M*	0.040***	-6.549***	-4.857***	1.082	0.028***	-3.924***	-4.117***	-0.652	0.006	-	-	-
	F	0.017**	-2.696*	-5.049***	-2.369	0.034***	-4.045***	-4.202***	-0.236	0.010*	-1.248	-0.265	1.034
CABIN	*R*	0.113***	-8.457***	-4.489***	4.414***	0.002	-	-	-	0.001	-	-	-
	I	0.069***	-4.750***	-0.844	3.343**	0.007	-	-	-	0.000	-	-	-
	A	0.102***	-6.430***	-2.322	4.460***	0.005	-	-	-	0.000	-	-	-
	S	0.100***	-7.033***	-2.768*	4.443***	0.003	-	-	-	0.002	-	-	-
	E	0.086***	-5.864***	-2.008	4.208***	0.010*	2.988**	-0.076	-3,366**	0.001	-	-	-
	C	0.086***	-7.655***	-3.437**	4.208***	0.008*	2.364	-0.097	-2.691*	0.000	-	-	-
ICB	*O*	0.002	-	-	-	0.009	-	-	-	0.001	-	-	-
ECR-R	*Anx.*	0.118***	7.816***	-4.777***	-11.632***	0.084***	-5.398***	-7.303***	-2.201	0.026***	5.028***	0.188	-5.203***
	Avo.	0.092***	5.831***	-3.704***	-9.093***	0.063***	-4.885***	-6.101***	-1.508	0.011*	2.851*	-0.435	-3.318**
MFQ-FF	*S. C*	0.047***	-1.615	4.056***	5.173***	0.287***	0.000	16.127***	15.522***	0.082***	3.693***	8.507***	6.373***
	H	0.061***	-2.231	4.362***	6.138***	0.291***	0.035	16.804***	15.046***	0.097***	2.948*	8.370***	6.862***
	I	0.057***	-1.720	4.477***	5.705***	0.297***	-0.598	16.362***	14.885***	0.081***	3.663***	8.197***	6.122***
	R	0.043***	-1.997	3.964***	5.378***	0.290***	-0.867	15.925***	14.265***	0.093***	3.992***	8.035***	5.668***
	S-V	0.050***	-1.971	3.964***	5.378***	0.281***	-0.376	16.498***	15.681***	0.086***	3.600**	7.893***	6.231***
	E	0.054***	-1.463	4.436***	5.380***	0.297***	-0.915	16.772***	14.829***	0.099***	3.942***	8.556***	6.431***
GSE	*O*	0.007	-	-	-	0.026***	-9.899***	-8.772***	-1.201	0.002	-	-	-
LOT-R	*O*	0.001	-	-	-	0.032***	-2.632*	-3.901***	-1.519	0.005	-	-	-
LMS	*R*	0.326***	-13.090***	-17.932***	- 6.741***	0.347***	-10.005***	-16.731***	-10.318***	0.022***	-2.525*	-0.280	2.572*
	M	0.296***	-10.344***	-16.477***	-8.042***	0.316***	-9.997***	-16.519***	-9.207***	0.046***	-5.166***	-0.930	3.612**
	I	0.273***	-10.506***	-17.160***	-8.09***	0.307***	-9.934***	-16.247***	-8.458***	0.025***	-4.260***	-0.930	3.612
EIS	*O*	0.132***	-11.259***	-10.156***	0.715	0.073***	-6.740***	-5.822***	-0.069	0.035***	-2.487*	-3.016**	-0.726
WLEIS	*S*	0.032***	-5.057***	-2.391	2.318	0.112***	-5.533***	-9.040***	-5.280***	0.010*	-3.275**	-1.541	1.827
	O	0.003	-	-	-	0.097***	-3.416**	-8.549***	-6.471***	0.019***	-3.091**	-2.129	-0.980
	U	0.012**	-3.645***	-2.977**	-0.237	0.067***	-3.995***	-8.205***	-4.470***	0.029***	-3.660***	-2.096	1.585
	R	0.009*	-1.918	-2.254	-0.537	0.061***	-3.024**	-7.754***	-5.017***	0.021***	-2.506*	-1.269	1.310
Empathy	*O*	0.109***	-8.777***	-3.338**	6.611***	0.038***	-3.843***	-5.828***	-2.417*	0.003	-	-	-

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

Table 10: Statistical test results for RAG condition with high-sensitivity personas (RQ1). 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.

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 11: Top 10 topics discovered per family (RQ2). Italicized and underlined words indicate 'as an AI' and first-person pronouns.

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 12: Starting sentence of a representative example, for each topic of parameter size groups (RQ2)

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 13: Starting sentence of a representative example, for each topic of GPT, LLaMA, and Mixtral (RQ2)

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 14: Starting sentence of a representative example, for each topic of Qwen (RQ2)

	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 15: Result of statistical tests for GPT3.5-turbo and GPT4o (RQ2). 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 16: Result of statistical tests for LLaMA3.1 model family (RQ2). 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 17: Result of statistical tests for Mixtral model family (RQ2). 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 18: Result of statistical tests for Qwen2 model family (RQ2). 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.