Can LLM Agents Maintain a Persona in Discourse?

Pranav Bhandari[a,b] [*], **Nicolas Fay**[c], **Michael Wise**[b], **Amitava Datta**[b], **Stephanie Meek**[d], **Usman Naseem**[e], **Mehwish Nasim**[a,b]
[a]Network Analysis and Social Influence Modelling (NASIM) Lab
[b]School of Physics, Mathematics and Computing, The University of Western Australia
[c]School of Psychological Sciences, The University of Western Australia
[d]School of Business and Law, Edith Cowan University, Australia
[e]School of Computing, Macquarie University, Australia

Abstract

Large Language Models (LLMs) are widely used as conversational agents, exploiting their capabilities in various sectors such as education, law, medicine, and more. However, LLMs are often subjected to context-shifting behaviour, resulting in a lack of consistent and interpretable personality-aligned interactions. Adherence to psychological traits lacks comprehensive analysis, especially in the case of dyadic (pairwise) conversations. We examine this challenge from two viewpoints, initially using two conversation agents to generate a discourse on a certain topic with an assigned personality from the OCEAN framework (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) as High/Low for each trait. This is followed by using multiple judge agents to infer the original traits assigned to explore prediction consistency, inter-model agreement, and alignment with the assigned personality. Our findings indicate that while LLMs can be guided toward personality-driven dialogue, their ability to maintain personality traits varies significantly depending on the combination of models and discourse settings. These inconsistencies emphasise the challenges in achieving stable and interpretable personality-aligned interactions in LLMs.

1 Introduction

Large language models (LLMs) have evolved from task solvers and general-purpose chatbots to sophisticated conversational agents capable of embodying distinct personas. This shift towards personalised agents, driven by LLMs' capacity for perception, planning, generalisation, and learning (Xi et al., 2025), has enabled context-sensitive discourse and opened up new possibilities across diverse domains. Persona, defined as conditioning AI models to adopt specific roles and characteristics (Li et al., 2024a), is a key element in this evolution.

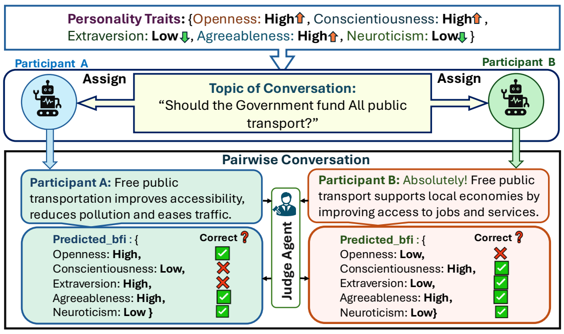


Figure 1: An example of inducing personality in LLM agents, followed by a discourse. A judge agent evaluates whether personality traits were adhered to within the discourse.

Personalised agents show promise in areas such as emotional support, training, and social skills development (Dan et al., 2024), and are increasingly explored for applications ranging from social science research (Zhu et al., 2025) to mimicking human behaviour (Jiang et al., 2023). Although there are various personalisation approaches, the incorporation of personas has been proven to be particularly effective in generating contextually appropriate responses and improving overall performance (Tseng et al., 2024; Dan et al., 2024).

Understanding how LLMs express and sustain personality traits in dynamic conversations is crucial, despite their tendency to generate neutral, balanced content. Existing work has explored personality in text using tools like the Big Five Inventory (BFI) (John et al., 1991) to infer and analyse personality profiles (Bhandari et al., 2025). However, two key gaps remain. First, it is unclear how consistently LLMs portray assigned personality traits during extended interactions, particularly in pairwise (dyadic) conversations where context shifts and adaptation are necessary. Second, robust methods are needed to evaluate the alignment between the expressed traits in the generated text and the intended psychological profile. We present an ex-

[*]Corresponding author: pranav.bhandari@uwa.edu.au

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29202–29218
November 4-9, 2025

ample in Figure 1.

While previous studies (Jiang et al., 2023; Kim et al., 2025) have made progress in demonstrating that LLMs can reflect assigned personality traits (often through personality questionnaires), a critical gap remains in understanding how consistently these traits are maintained in generated content, particularly within dynamic conversational settings. Although assigning personality traits to conversational agents often yields positive results in controlled settings, this does not guarantee that the generated content effectively expresses those traits, nor does it quantify the degree of expression. Our work differs in two key ways: **(1)** we study trait adherence in pairwise conversational settings where agents must respond contextually while maintaining personality, and **(2)** we incorporate a multi-judge evaluation framework to assess both trait detectability and inter-rater reliability, offering a more comprehensive view of personality alignment in LLMs discourse.

This work aims to investigate how effectively LLMs express assigned personality traits in generated dialogue. Specifically, we explore whether and how LLMs maintain Big Five Personality traits, which are represented as the **OCEAN** framework (Husain et al., 2025) (*Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism*), during dyadic conversations. We employ a novel agent-based evaluation framework where two LLM agents, each assigned a distinct OCEAN personality profile, engage in a conversation on a given topic. Subsequently, independent LLM agents (*judges*) assess the generated dialogue to determine the consistency between expressed and assigned traits. This approach allows us to analyse not only whether LLMs reflect personality, but also the peculiarities in trait expression and the challenges of maintaining personality consistency within dynamic conversational contexts.

This work seeks to address the following research questions:

RQ1: Are all OCEAN traits equally prominent in generated conversations?

RQ2: How accurately do LLMs as *judge agents* predict the assigned traits from discourse?

RQ3: Do different LLM judge agents consistently extract the assigned personality traits from the conversations?

2 Related Work

While personality traits are central to the human-like behaviours exhibited by LLMs, the systematic psychological evaluation of these traits remains largely underexplored and warrants further investigation (Zhu et al., 2025). The recent literature has looked at designing (Klinkert et al., 2024), improving (Huang et al., 2024), investigating (Frisch and Giulianelli, 2024; Zhu et al., 2025), customising (Han et al., 2024; Dan et al., 2024; Zhang et al., 2018) and exploring (Zhu et al., 2025; Han et al., 2024) personality traits. The scope of our work lies both in generating and extracting personality traits embedded within discourse.

Han et al. (2024) contribute towards the generation of synthetic dialogues through LLMs. A five-step generation process is used where personality is induced through personality character. Special consideration on prompts is made to infer Pre-trained Language Models (PLM) in generating dialogues. This is because dialogue generation is a challenging task, especially with many constraints and maintaining personality traits. Unlike traditional methods of curating datasets by humans, the authors leverage the capability of PLM to generate synthetic data that is easily scalable. The use of these synthetic datasets significantly improved the ability of LLMs to generate content that is more tailored towards personality traits. While the research is broad, its dataset is limited to Korean and focuses on a single personality trait, which may hinder balanced trait prediction.

Although designing and customising personality traits in LLM is an active area of research, this work focusses on inducing and evaluating these traits through discourse generation (Yeo et al., 2025). Jiang et al. (2023) study LLMs' ability to express personality traits in essay generation using both human and LLM-based evaluations. They apply LIWC analysis and human annotation to assess GPT-generated content, finding positive correlations with intended traits. However, their work is limited to closed models, single-ended generation, and a small output set. Non-GPT models were excluded due to inconsistent instruction-following. To address these gaps, we adopt structured prompting and extend the analysis to multi-turn dialogues across diverse models, enabling broader evaluation of trait expression in dynamic settings.

Sun et al. (2024) argues that personality detection should be evidence-based rather than a clas-

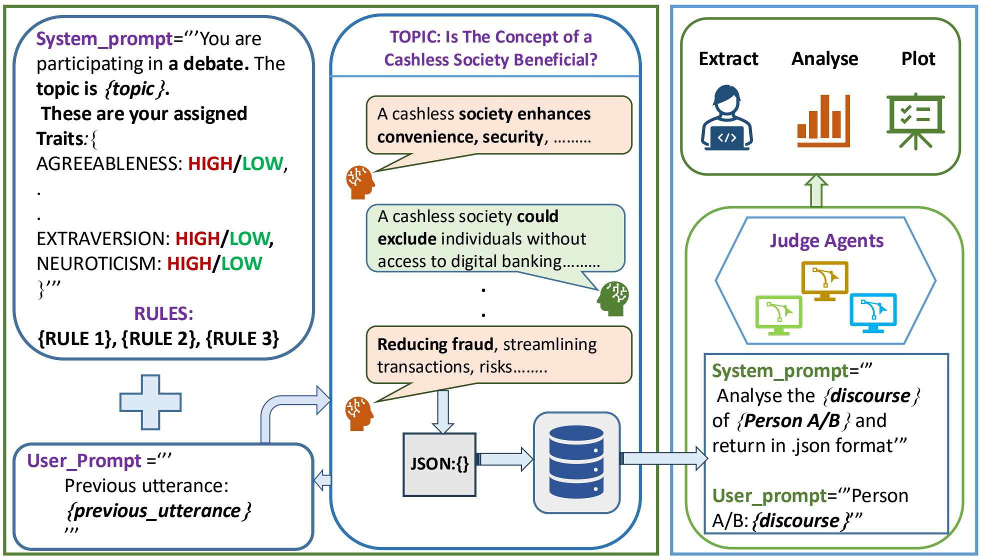


Figure 2: Methodology of the paper. The **System prompt** induces both the personality traits and the topic of discourse, while the **User prompt** provides the conversational context. Importantly, the context includes the entire dialogue history (rather than just the most recent utterance), which is then used to generate the next response to ensure the context is relevant and coherent. The resulting dyadic conversations are subsequently analysed by **Judge Agents** to report the findings.

sification task, enhancing explainability. They introduce the Chain of Personality Evidence (CoPE) dataset for personality recognition in dialogues, addressing state and trait recognition. However, limitations include model specialisation and the availability of a small dataset in Chinese, leaving gaps in the personality trait recognition research.

Prompting methods: Different methods for assigning personality traits are used in the literature, mainly categorising explicit or implicit mention of personality traits or training-based methods. Most studies focus on implementing the OCEAN models for the agents (Bhandari et al., 2025). One common way of assigning personality traits is through direct allocation of personalities and assigning the personality traits to the agents (Jiang et al., 2023). Another commonly followed methodology is passing content that infers the traits but does not directly mention them (Sun et al., 2024; Han et al., 2024). Personality is also assigned through fine-tuning where distinct fine-tuned models represent distinct personalities. We believe that providing clear instructions about the personas would clear the ambiguity and hence prompt the use of the direct allocation method.

Evaluation: LLMs are increasingly used to evaluate personality traits from the text. While their accuracy is still under study, they offer a cost-effective and efficient approach.

Zhu et al. (2025) use closed-source models (GPT-4o and GPT-4o-mini) to infer the BFI traits and extract the scores.

Authors present the findings that the effectiveness of LLMs in predicting personality traits increased as they were prompted with an intermediate step of BFI-10 (Rammstedt, 2007) questionnaires. Two main metrics were used to benchmark the ability of LLMs: correlation and mean difference, where correlation measured the ability to capture structural relationships and mean difference captured absolute prediction accuracy. We also adapt these metrics to evaluate the content produced by LLMs in our agent ecosystem. Different validation datasets relating to personality traits include: Essay Dataset (Yeo et al., 2025), myPersonality (Zhu et al., 2024), and Twitter Dataset (Shu et al., 2024).

In summary, the main problems identified in the literature are the use of closed-source models, the lack of analysis in content generation consisting of context-shifting behaviour, and the lack of use of standard evaluation metrics. Furthermore, one of the main challenges in incorporating personality traits is understanding whether all five traits are effectively adhered to in the content that is produced. We aim to address some of these problems through this research.

3 Methodology

We present the methodology of this work in Figure 2. In an agent-based setting the methodology is operationalised in 4 phases: ***P**ersonifying agents*, ***G**enerating discourse*, ***E**xtracting personality within discourse*, and ***E**valuation*. A detailed explanation of the modular approach is presented in subsequent sections.

We adopted an iterative approach to refine the methodology. At first, producing coherent discourse between the models was difficult because of synchronization issues, over-generalisation, repetition of prompts, and the models explicitly mentioning the assigned personality. To address these, we progressively refined the prompts and adjusted the setup until the dialogues became more natural and consistent. Because dyadic conversations are highly dependent on earlier turns, even a single unjustified or low-quality response could derail the entire interaction. Therefore, we placed special emphasis on designing prompts and controls that encouraged complete and sensible conversations. Finally, to ensure diversity, we performed a similarity check across all dyadic conversations to confirm that the models were not reproducing identical dialogues.

We selected GPT models from OpenAI(OpenAI, 2024) and LLaMA models from Meta(Patterson et al., 2022) due to their popularity and reach. As the landscape rapidly evolved, we expanded our scope to include DeepSeek[1] to ensure broader coverage and comparison across architectures.

Since the generation of essays on a particular topic has been explored in literature such as (Kim et al., 2025; Yeo et al., 2025), we wanted to explore the generation of discourses, particularly for two reasons **1)** The complexity of the topic increases and maintaining a progressive discussion given the explicit persona is a difficult task. **2)** It is also interesting to understand the consistency in the personality during a conversation. We chose a structured dyadic debate format deliberately due to its ability to elicit trait-driven argumentative behaviour. Debates provide a cognitively demanding and structured setting where traits such as Agreeableness, Conscientiousness, and Extraversion are more likely to manifest distinctly. This choice enables consistent topic framing while encouraging trait-relevant expression across contrasting viewpoints, which is essential for measuring sustained personality adherence.

Dataset: We have carefully selected 100 different topics incorporating various domains that require, *ethical, moral, social or political* considerations [2] and 20 different combinations of random traits (more in Appendix A).

3.1 Prompt formation

The prompt formation is an essential part of our methodology. Since other agents analyse the discourse and we draw the results based on the discourse, it must be structured robustly to ensure reliability and objective evaluation.

The *system and user prompts were re-initialised at each turn to ensure that assigned personality traits are consistently represented throughout the conversation*. This repetition helps reinforce trait adherence across all dialogue steps. Thus, the predefined traits are not used as a one-time instruction but as a continuously reinforced signal during generation, making them a valid basis for downstream evaluation metrics.

Prompting for LLMs is carried out through specific prompting methods where agents are assigned roles to convey requirements and expected outcomes. Usually, the *system and user* roles are passed as arguments (Yeo et al., 2025) in which the system role is responsible for defining the behaviour and limiting the scope of response. The user role is used for defining the input. Despite strict adherence to these techniques, agents may still be overwhelmed by excessive constraints.

System Prompt: The system prompt in our work contains the rules for debates carried out on a specific topic. Structured prompts enhance clarity for agents, improve effectiveness, and help users create inclusive prompts despite multiple constraints. Although the formatting of the prompts varies according to the model specifications, they contain the following information.

- The traits are assigned in two forms of extremities: *High or Low*.
- You are a participant in a discourse in which the topic is *topic* and presented with the following traits *traits*.
- Assigned personality traits must be maintained throughout the conversation but not explicitly mentioned in the utterances.
- Each utterance must be under 50 words and the previous utterance needs to be addressed.

[1] DeepSeek models

[2] Debate Topics

Our use of binary High/Low assignments is a deliberate simplification to *enable clearer interpretability, controlled trait induction, and tractable evaluation* using classification-based metrics (HTA/LTA). Although the OCEAN model is spectrum-based, in early pilot studies we found that using continuous trait values introduced ambiguity in both generation and evaluation, especially since LLMs are not inherently optimized for regression tasks (Tang et al., 2024). Binary prompts reduce this complexity and align with prior work in personality-aligned LLMs (Li et al., 2024b; Vu et al., 2024), where discrete trait definitions improved controllability and instruction follow-up. Moreover, spectrum-based trait representation demands more granular annotation and evaluation, which was not feasible due to time and resource constraints. Hence, we opt for a binary framework to ensure clarity, consistency, and reliable evaluation in this study.

User Prompt: User prompt in this case contributes to an important role in shaping the conversation because the previous discussions are passed through the user prompt to generate the next utterance.

3.2 Validation

Validation involves both human and agent-based evaluation. **(1)** A random sample of 60–70 discourses per category was manually assessed for *length*, *coherence*, *quality*, and *personality cue presence* (Appendix D), though traits were not manually labelled as High/Low. **(2)** Utterance similarity was measured to ensure diversity of arguments. LLMs have been shown to simulate and infer Big Five traits reliably (Jiang et al., 2023; Han et al., 2024; Zhu et al., 2025), particularly with structured prompting. Our multi-agent judge setup is supported by (Frisch and Giulianelli, 2024), who emphasise inter-model agreement as critical for robust trait attribution. To assess the validity of our judge agents, we evaluated GPT-4o on 1,000 essays from the validated Essay dataset(Mairesse et al., 2007) with known Big Five labels (more in Appendix E).

4 Evaluation

Once the discourses are generated, each is evaluated by *Judge agents*, which return trait predictions for each speaker in *JSON* format. To reduce bias toward agent-generated content, we specify in the prompt that the utterances are "human-generated." All judge models were run using deterministic decoding settings (temperature = 0) to ensure consistent and replicable outputs across evaluations. The following evaluations are performed:

4.1 Discourse alignment with Assigned Personality Traits

The discourse alignment with assigned personality traits is an important part of this analysis as it depicts if the personality traits are reflected in the contents generated by the agents. We analyse if the discourses linguistically align with the assigned personality traits. Various factors like language, tone and argument structures contribute towards the alignment of personality traits with the content produced (Pennebaker and King, 1999).

Linguistic Inquiry and Word Count (LIWC-22) (Boyd et al., 2022) analysis is a widely used tool for this category that classifies words into psychological and linguistic categories. Its relevance to the Big Five Analysis has been established in both psychological studies and computational models (Jiang et al., 2023; Mairesse et al., 2007). Previous research explains how natural language and linguistic markers can effectively serve as indicators of personality traits (Ireland and Mehl, 2014). For instance, extroverts tend to use more positive words and social process words to reflect their sociable nature. Linguistic markers are successfully able to understand and predict the personality traits in given text (Mairesse et al., 2007). We use the capabilities of LIWC-22 to extract the linguistic features and systematically map the five personality traits from the data to analyse the results. While LIWC-22 is a static lexical analysis tool and may not fully capture the evolving nature of personality across conversational turns, our methodology compensates for this by reinitializing personality-conditioned prompts at each utterance, thereby encouraging trait-relevant expression that LIWC captures at the utterance level. To clarify further, both the LIWC-22 analysis and the judge models' personality trait predictions are performed at the full conversation level, not on individual utterances. This approach was chosen because personality traits are more reliably expressed and interpreted over a sequence of dialogues, rather than in isolated turns.

4.2 Personality prediction by Judge Agents

With access to both the assigned traits (Section 3.1) and inferred traits (Section 4) using different judge agents, we begin by calculating the accuracy of

Judge	Comparison	Personality Traits	HTA_P1	LTA_P1	HTA_P2	LTA_P2
GPT-4o	GPT-4o vs GPT-4o-mini	Ag	97.1	57.1	99.7	38.9
		Op	98.4	30.9	99.5	15.0
		Co	97.1	12.8	97.6	11.0
		Ex	63.1	92.2	63.2	86.7
		Ne	31.4	96.6	28.9	97.0
	GPT-4o vs LLaMA-3.3-70B-Instruct	Ag	91.2	73.0	98.3	47.3
		Op	95.1	37.6	96.0	21.6
		Co	96.3	24.0	97.8	7.8
		Ex	43.9	95.8	43.3	94.7
		Ne	62.1	86.0	25.7	94.3
	GPT-4o vs DeepSeek	Ag	91.8	18.6	99.1	40.0
		Op	99.1	6.4	97.7	11.0
		Co	97.7	9.9	99.2	2.4
		Ex	7.4	99.3	1.2	100.0
		Ne	10.3	96.0	11.3	95.0
GPT-4o-mini	GPT-4o vs GPT-4o-mini	Ag	97.8	49.0	99.3	34.4
		Op	98.4	26.5	99.5	13.0
		Co	97.3	15.4	97.0	18.2
		Ex	66.8	83.0	70.4	76.1
		Ne	14.3	99.2	11.4	99.4
	GPT-4o vs LLaMA-3.3-70B-Instruct	Ag	95.1	59.5	98.8	47.2
		Op	97.5	28.0	97.1	20.1
		Co	95.4	30.4	97.0	12.3
		Ex	54.2	89.3	49.3	88.7
		Ne	33.8	98.5	13.4	99.3
	GPT-4o vs DeepSeek	Ag	92.2	17.1	98.8	38.7
		Op	99.9	7.6	97.2	14.1
		Co	97.6	9.7	97.9	5.2
		Ex	14.7	97.3	10.3	97.4
		Ne	4.1	99.2	2.1	99.2
LLaMA	GPT-4o vs GPT-4o-mini	Ag	98.1	48.1	99.9	33.3
		Op	94.5	38.6	97.4	21.7
		Co	98.2	9.3	98.8	7.7
		Ex	64.5	77.7	67.6	74.5
		Ne	20.6	96.6	20.7	97.0
	GPT-4o vs LLaMA-3.3-70B-Instruct	Ag	95.2	61.4	99.4	40.2
		Op	93.4	40.8	92.2	27.7
		Co	98.1	14.9	99.1	4.7
		Ex	54.0	82.8	47.3	84.2
		Ne	49.2	92.3	18.9	96.3
	GPT-4o vs DeepSeek	Ag	95.9	13.2	99.8	32.2
		Op	98.3	10.9	93.6	21.6
		Co	96.6	10.6	98.9	2.2
		Ex	24.9	90.0	11.4	95.0
		Ne	7.3	96.5	9.3	95.5
Qwen	GPT-4o vs GPT-4o-mini	Ag	95.0	43.2	97.7	35.7
		Op	76.8	60.6	80.3	53.3
		Co	67.2	55.1	64.1	57.6
		Ex	15.6	94.9	12.6	95.0
		Ne	27.7	90.8	28.6	91.9
	GPT-4o vs LLaMA-3.3-70B-Instruct	Ag	91.9	49.7	95.0	36.6
		Op	77.9	57.5	70.4	57.7
		Co	69.5	57.1	75.7	40.8
		Ex	11.9	95.1	16.1	94.5
		Ne	49.2	88.9	23.3	90.9
	GPT-4o vs DeepSeek	Ag	81.0	25.8	92.6	31.5
		Op	80.0	32.2	60.7	55.2
		Co	74.7	35.3	85.1	23.7
		Ex	6.5	95.9	1.9	97.9
		Ne	11.2	92.3	12.5	94.0

Metrics(%)

Table 1: Calculation of High Trait Classification Accuracy(**HTA**) and Low Trait Classification Accuracy(**LTA**) for **Participants 1 and 2** across all the conversations for all the **Judge Agents** (GPT-4o-mini in Appendix.).

the models' predictions (a.k.a. inferred traits). We calculate the accuracy of prediction in two different ways: the accuracy of predicted *High* for each trait as High Trait Classification Accuracy(HTA) and finally accuracy of predicted *Low* for each trait as Low Trait Classification Accuracy(LTA). Recall, that we assign a high or a low value for each *OCEAN* trait while assigning personalities in Section 3.1. We create a confusion matrix for this labelling all the True and False predictions of High and Low values to compute the HTA and LTA values.

HTA measures how well the models classify traits assigned as High originally. This is computed by creating a confusion matrix for correct and incorrect classifications. HTA is calculated by dividing the total correctly classified High by the total number of High cases.

LTA on the other hand measures how well the models classify traits assigned as Low originally. It is calculated by dividing the total correctly classified Low by the total number of Low cases. An important aspect of this study is understanding potential bias in classification into High or Low traits. While overall accuracy may be high, we focus on whether both categories are proportionately represented.

4.3 Inter-rater reliability among the models

Inter-rater reliability is the measure to understand the agreement between the models. Kappa statistics(κ) is a common method to assess the consistency of ratings among raters (Judge LLMs) (Pérez et al., 2020).

We computed `Fleiss' Kappa` by first gathering personality trait predictions from five different judge models. Each model analysed debates across multiple topics and rated Big Five personality traits for two participants (P1 & P2). We structured the data so that all model ratings for the same Topic-Trait pair were aligned, ensuring consistency in comparison. After validation, we reformatted the dataset into a matrix where each row represented a topic-trait combination. The matrix contained counts of how many models classified the trait as *High* or *Low* for both P1 and P2 separately. We calculated the inter-model agreement for each trait using Python's 'statsmodels'[3] package, specifically the `Fleiss' Kappa` function to extract the consistency of various judge models across all topics.

While the first measure explores the accuracy with which the models correctly identify *High* and *Low*, respective to the ground values, this method explores the agreement between the models for a particular trait at a time, irrespective of the base values.

5 Results

Four models are involved in the creation of discourse in different combinations (GPT-4o vs. GPT-4o-mini, GPT-4o vs. Llama-3.3-70B-Instruct, GPT-4o vs. Deepseek-llm-67B-Chat). All of these models have been set up at higher temperatures (>0.8) to allow creativity during discourse generation. Limited by resources(NVIDIA A6000 GPU), the larger models such as Llama-3.3-70B-Instruct and Deepseek-llm-67B-Chat, were quantized to generate discourse. Models were given varying max token limits (OpenAI: 150, LLaMA: 200, DeepSeek: 350) to accommodate generation behaviour and prevent incomplete responses. Furthermore, to contain the noise produced by *thinking* tokens by DeepSeek reasoning models, the *Chat* model was used instead of the *Reasoning* Model.

For the evaluations of the generated discourse, we used five different models: GPT-4o, GPT-4o-mini, Llama-3.3-70B-Instruct, Qwen-2.5-14B-Instruct-1M, and Deepseek-llm-67B-Chat — *the judge agents*. The idea is to include a variety of models(both small and large) and understand the consistency in the results.

While GPT models adhered to instructions reliably with minimal prompting, LLaMA and DeepSeek required additional filtering due to prompt repetition and formatting issues; details on prompt adaptations are provided in the Appendix B.2.

5.1 Discourse Alignment with assigned personality traits

Figure 3 presents the accuracy of personality trait depiction for Participants 1 and 2, measured using LIWC-22. GPT-4o-mini achieved the highest accuracy for Agreeableness across all dialogues. However, GPT-4o's Agreeableness accuracy decreased substantially (from 68% and 65% to 52%) when conversing with Deepseek than GPT-4o-mini and Llama-3.3, suggesting a potential shift in personality expression depending on the interlocutor, similar to human behaviour (Atherton et al., 2022).

Openness was the trait least accurately represented in all dialogues, with a maximum accuracy of 51%. This suggests that expressing Openness is particularly challenging for these LLMs. Llama-3.3 exhibited the highest Conscientiousness, while GPT-4o showed the highest Extraversion. However, these differences were not statistically significant, and trait expression varied depending on the conversational partner. GPT-4o's Neuroticism depiction was most accurate when interacting with Llama-3.3. This variability in traits and conversational settings directly addresses **RQ1**, confirming that all OCEAN traits are not equally prominent in generated conversations.

When comparing pairwise dialogues, GPT-4o vs. GPT-4o-mini and GPT-4o vs. Llama-3.3 showed similar performance. However, GPT-4o vs. Deepseek dialogues exhibited significantly different results. We observed that Deepseek struggled to consistently follow instructions from the prompts (even though the prompts were minimally adapted across models). Deepseek's generated text was also the most inconsistent in length compared to other models, which may have contributed to the observed differences.

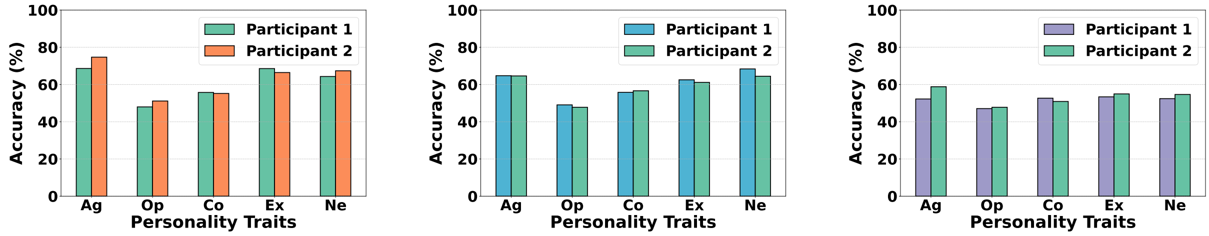
Trait	Discourse 1		Discourse 2		Discourse 3	
	P1	P2	P1	P2	P1	P2
Agr	0.500	0.557	0.242	0.692	0.518	0.532
Ope	0.699	0.420	0.534	0.631	0.250	0.430
Con	0.352	0.366	0.502	0.421	0.330	0.367
Ext	0.123	0.097	0.235	0.105	0.287	0.260
Neu	0.480	0.293	0.233	0.463	0.351	0.389

Table 2: Fleiss' Kappa Scores for Personality Trait Agreement. *Discourse 1:* **GPT-4o vs. GPT-4o-mini**, *Discourse 2*: **GPT-4o vs. Llama-3.3-70B-Instruct** and *Discourse 3*: **GPT-4o vs. Deepseek-llm-67b-chat**. P1 and P2: Participants 1 and 2 respectively.

5.2 Personality Prediction by Judge Agents

Table 1 presents personality prediction results from each judge model. We observed several notable patterns across traits, judge models, and conversation types.

[3] statsmodels



(a) GPT-4o vs. GPT-4o-mini (b) GPT-4o vs. LLaMA-3.3 (c) GPT-4o vs. DeepSeek

Figure 3: This figure depicts the accuracy of conveying the assigned personality traits by Participants 1 and 2 using the LIWC analysis. LIWC-22 analysis is performed at the full conversation level, rather than on individual utterances.

Analysis Across Judge Models: For Agreeableness, Openness, and Conscientiousness, GPT-4o, GPT-4o-mini, and LLaMA-3.3-70B-Instruct yield high prediction accuracy (>90%) for both participants. In contrast, Qwen-2.5-14B-1M underperforms for Openness and Conscientiousness but maintains reasonable scores for Agreeableness. Larger models (e.g., GPT-4o, LLaMA-3.3) are more effective at predicting *High* trait values, while Qwen-2.5 achieves better accuracy for *Low* Openness and Conscientiousness. Overall, High trait values are more accurately predicted than Low values for these three traits across most models.

Trait-Specific Patterns: For Extraversion and Neuroticism, High values are generally underpredicted across all judge models and discourses. High Neuroticism is particularly difficult to detect, potentially due to LLMs' avoidance of highly negative or anxious content. However, GPT-4o performs better than others, achieving 62% precision for High Neuroticism in some settings. Notably, judge performance on High Neuroticism is weakest in GPT-4o vs. DeepSeek conversations. Due to over 40% invalid responses, DeepSeek was excluded as a judge model in Table 1.

Analysis Across Conversations: Judge accuracy for High Neuroticism and Extraversion was notably lower in GPT-4o vs. DeepSeek conversations compared to the other two (GPT-4o vs. GPT-4o-mini and GPT-4o vs. LLaMA-3.3). This suggests that trait expression in discourse is more difficult to detect when involving models with inconsistent persona adherence or generation stability. However, judges showed consistency in trait evaluation among participants: If one judge rated a participant high on Agreeability, other judges often did the same.

RQ2 and RQ1: These results address **RQ2**, showing that LLMs can conditionally predict assigned traits from dialogue, with performance varying by trait and trait polarity (High/Low). The uneven predictability across traits, particularly the difficulty with High Neuroticism and Extraversion, partly addresses **RQ1**, implying sensitivity biases in judge agents.

5.3 Inter Model Agreement

Table 2 presents the Fleiss' Kappa statistics, measuring inter-model agreement on personality trait judgments for Participants 1 and 2 across all dialogues.

In Discourse 1, Agreeableness showed moderate agreement ($\kappa > 0.5$) for both participants. Openness agreement was substantial for Participant 1 but moderate for Participant 2. Conscientiousness and Neuroticism exhibited fair to moderate agreement. Notably, Extraversion showed the lowest agreement, indicating poor reliability in its assessment.

Discourse 2 revealed minimal Agreeableness agreement for Participant 1 but substantially higher agreement for Participant 2, highlighting fluctuations in judging this trait. Openness maintained moderate to substantial agreement. Conscientiousness and Extraversion agreement increased compared to Discourse 1, though Extraversion remained low overall. Neuroticism agreement showed a reversed trend, with lower agreement for Participant 1 and higher for Participant 2.

In Discourse 3, Agreeableness agreement remained moderate. Openness agreement decreased drastically. Conscientiousness, Extraversion, and Neuroticism agreement was stable between participants but only slight to fair.

These results address **RQ3**, demonstrating inconsistent inter-model agreement on personality traits. Agreeableness and Openness agreement fluctuated across dialogues. The consistently low Extraversion agreement indicates significant challenges in its reliable assessment. This variability underscores the non-uniformity of personality alignment in LLMs, highlighting difficulties in achieving stable and interpretable personality-driven interactions.

6 Conclusion

This paper provides a comprehensive evaluation of trait adherence in LLM agents engaged in dyadic conversations. Our findings highlight the significant challenges in achieving consistent and interpretable personality-aligned interactions. While LLMs can be guided to exhibit certain personality traits, their ability to maintain these traits across dynamic conversations varies considerably. Future work should explore more sophisticated methods for instilling and evaluating personality, investigating the impact of dialogue context and developing metrics for assessing the nuances of personality expression in LLMs. Exploring fine-tuning strategies or reinforcement learning approaches for improving consistency would also be valuable.

Limitations

While human-annotated trait labels would provide a valuable benchmark for validating model performance, obtaining such annotations at scale requires significant time and domain expertise, which was beyond the scope of this study. To ensure clarity and interpretability, we simplified the trait representation using binary values, allowing judge agents to operate under a well-defined classification framework. Although this design choice makes the analysis tractable, we acknowledge that personality traits naturally lie on a continuous spectrum, and regression-based trait inference would be a promising extension for future work.

Another limitation stems from the lack of comparable prior work with a similar conversational agent setup, which makes traditional baseline comparisons non-trivial. To mitigate this, we included a diverse pool of judge agents, measured inter-model agreement, and incorporated a zero-shot generation reference as a soft baseline to assess the effectiveness of personality prompting. Still, there is a risk of bias, as language models may incorporate their own implicit judgments into discussions, potentially influencing personality assessments. We also note that LIWC is used only as a heuristic indicator rather than a definitive ground truth.

Further, the length of dyadic conversations presents another challenge, as there is no widely accepted standard for how long a dialogue should be to ensure a reliable evaluation. This uncertainty raises questions about whether longer or shorter exchanges might yield different insights, adding a layer of complexity to the interpretation of results. Finally, our reliance on debate-style conversations may limit the generalizability of findings to other dialogue types such as casual or task-oriented exchanges.

Ethical Considerations

We do not collect any personal information and views for the creation of the discourse dataset or refer to any kind of personal traits from any sources to judge the nature of conversations. All the discourses are created by LLM agents. Topics provided for discussion for the agents are debatable but do not involve or promote the thought of violence, hatred or extremism of any kind to anyone.

We use open and closed-source models that are available off-the-shelf and accessible to the general public. No changes in the model architecture have been made. Some hyperparameters have been adjusted to meet our expectations of the results, but they have been mentioned clearly in the paper. LLMs have the possibility of introducing bias in their results as per numerous studies. The dataset generated by the conversing agents has not been made public, but we do plan to publish it for further studies with careful ethical consideration and approvals. The results do present bias in predicting the BFI from the discourses but are solely limited to LLMs as judges.

The content of LLM agents is subject to change if they are altered, fine-tuned, and tempered in different ways, which is a potential risk.

References

Olivia E Atherton, Angelina R Sutin, Antonio Terracciano, and Richard W Robins. 2022. Stability and change in the big five personality traits: Findings from a longitudinal study of mexican-origin adults. *Journal of Personality and Social Psychology*, 122(2):337.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM Web Conference*.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10.

Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. 2024. P-tailor: Customizing personality traits for language models via mixture of specialized lora experts. *arXiv preprint arXiv:2406.12548*.

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, St. Julians, Malta. Association for Computational Linguistics.

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. Psydial: Personality-based synthetic dialogue generation using large language models. *arXiv preprint arXiv:2404.00930*.

Muhua Huang, Xijuan Zhang, Christopher Soto, and James Evans. 2024. Designing llm-agents with personalities: A psychometric approach. *arXiv preprint arXiv:2410.19238*.

Waqar Husain, Areen Jamal Haddad, Muhammad Ahmad Husain, Hadeel Ghazzawi, Khaled Trabelsi, Achraf Ammar, Zahra Saif, Amir Pakpour, and Haitham Jahrami. 2025. Reliability generalization meta-analysis of the internal consistency of the big five inventory (bfi) by comparing bfi (44 items) and bfi-2 (60 items) versions controlling for age, sex, language factors. *BMC psychology*, 13(1):20.

Molly E Ireland and Matthias R Mehl. 2014. Natural language use as a marker. *The Oxford handbook of language and social psychology*, pages 201–237.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology*.

Hongjin Kim, Jeonghyun Kang, and Harksoo Kim. 2025. Can large language models differentiate harmful from argumentative essays? steps toward ethical essay scoring. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8121–8147.

Lawrence J Klinkert, Steph Buongiorno, and Corey Clark. 2024. Evaluating the efficacy of llms to emulate realistic human personalities. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 20, pages 65–75.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024a. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7283–7298.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2024b. Big5-chat: Shaping llm personalities through training on human-grounded data. *arXiv preprint arXiv:2410.16491*.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Preprint*, arXiv:2204.05149.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Jorge Pérez, Jessica Díaz, Javier Garcia-Martin, and Bernardo Tabuenca. 2020. Systematic literature reviews in software engineering—enhancement of the study selection process using cohen's kappa statistic. *Journal of Systems and Software*, 168:110657.

Beatrice Rammstedt. 2007. The 10-item big five inventory. *European Journal of Psychological Assessment*, 23(3):193–201.

Zhiyao Shu, Xiangguo Sun, and Hong Cheng. 2024. When llm meets hypergraph: A sociological analysis on personality via online social networks. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2087–2096.

Lei Sun, Jinming Zhao, and Qin Jin. 2024. Revealing personality traits: A new benchmark dataset for explainable personality recognition on dialogues. *arXiv preprint arXiv:2409.19723*.

Eric Tang, Bangding Yang, and Xingyou Song. 2024. Understanding llm embeddings for regression. *arXiv preprint arXiv:2411.14708*.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Huy Vu, Huy Anh Nguyen, Adithya V Ganesan, Swanie Juhng, Oscar NE Kjell, Joao Sedoc, Margaret L Kern, Ryan L Boyd, Lyle Ungar, H Andrew Schwartz, et al. 2024. Psychadapter: Adapting llm transformers to reflect traits, personality and mental health. *arXiv preprint arXiv:2412.16882*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Haein Yeo, Taehyeong Noh, Seungwan Jin, and Kyungsik Han. 2025. PADO: Personality-induced multi-agents for detecting OCEAN in human-generated texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5719–5736, Abu Dhabi, UAE. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Jianfeng Zhu, Ruoming Jin, and Karin G Coifman. 2025. Investigating large language models in inferring personality traits from user conversations. *arXiv preprint arXiv:2501.07532*.

Yangfu Zhu, Yue Xia, Meiling Li, Tingting Zhang, and Bin Wu. 2024. Data augmented graph neural networks for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 664–672.

A Sample of Topics and Trait Combinations Used

Samples of *topics* used for debate:

```
"Is the concept of a universal language
    beneficial?",
"Should the government regulate the
    pharmaceutical industry?",
"Is the use of nuclear energy
    justified?",
"Should the government provide free
    public transportation?",
"Is the concept of a cashless society
    beneficial?",
"Should the government regulate the
    gaming industry?"
```

Trait combinations samples to assign personas to Agents:

```
{"Agreeableness": "High", "Openness":
    "Low", "Conscientiousness": "High",
    "Extraversion": "Low",
    "Neuroticism": "High"},
{"Agreeableness": "Low", "Openness":
    "High", "Conscientiousness": "Low",
    "Extraversion": "High",
    "Neuroticism": "Low"},
{"Agreeableness": "High", "Openness":
    "High", "Conscientiousness": "Low",
    "Extraversion": "High",
    "Neuroticism": "High"},
{"Agreeableness": "Low", "Openness":
    "Low", "Conscientiousness": "High",
    "Extraversion": "Low",
    "Neuroticism": "Low"},
{"Agreeableness": "High", "Openness":
    "High", "Conscientiousness":
    "High", "Extraversion": "Low",
    "Neuroticism": "Low"}
```

B System and User prompts

We use different *System and User* prompts to extract the discourses and ratings from the conversing and judge agents. We experimented with various prompt formulations. Initially, we tested unstructured and minimal prompts to observe how well models understood the task. OpenAI models performed reliably even with loosely defined instructions, while LLaMA and DeepSeek required more structured prompts to produce coherent and persona-aligned responses. Once a stable structure was established, we varied the wording while keeping the meaning intact and found that performance remained consistent. This suggests that, given a clear prompt format, the models are robust to minor wording changes.

B.1 Discourse Generation

The *system prompt* to generate the discourses:

```
SYSTEM_PROMPT = ''' f"You are
    participating in a structured
    debate on: '{topic}'\n"
"Your responses should reflect these
    personality traits:\n"
f"- Agreeableness:
    {traits['Agreeableness']}\n"
f"- Openness: {traits['Openness']}\n"
f"- Conscientiousness:
    {traits['Conscientiousness']}\n"
f"- Extraversion:
    {traits['Extraversion']}\n"
f"- Neuroticism:
    {traits['Neuroticism']}\n\n"
"Rules:\n"
"- Maintain these personality traits
    (DO NOT EXPLICITLY MENTION IN TEXT)
    at all
times during your conversation\n"
"- Keep responses under 50 words\n"
"- Maintain your personality
    consistently\n"
"- Address previous arguments directly
    but do not repeat what
the other speaker said.\n"
"- End with proper punctuation" ''''
```

The *user prompt* carries the previous argument :

```
USER_PROMPT = """Previous
    Argument:f"{previous_arguement}" """
```

B.2 Adaptations of prompts required for Llama and Deepseek models

As discussed in the Results section, the prompts needed further improvements and details for the Llama and Deepseek for better results. While the USER_PROMPT was the same, the SYSTEM_PROMPT is adapted as follows:

```
SYSTEM_PROMPT =  f"<|start_header_id|>
    system <|end_header_id|> \n"
f"You are engaged in a formal debate
    about: '{topic}'\n\n""CRITICAL
    INSTRUCTIONS:\n"
"1. Your responses must NATURALLY
    reflect your assigned personality
    through:\n"
" - Word choice (e.g., cautious vs.
    bold language)\n"
" - Argument style (e.g.,
    detail-focused vs. big-picture)\n"
" - Emotional tone (e.g., enthusiastic
    vs. reserved)\n\n"
"2. Strict Response Requirements:\n"
" - 25-50 words EXACTLY \n"
" - Must end with proper punctuation\n"
" - Never reference being an AI or
    language model\n"
" - Never repeat the opponent's phrases
    verbatim\n\n"
"3. Debate Technique:\n""   - Directly
    engage with the previous point
    made\n"
" * Maintain personality through:\n"
" * How you structure arguments\n"
" * How you counter opponents\n\n"
"4. Prohibited behaviours:\n"
" - No disclaimers about neutrality\n"
" - No meta-commentary about the debate
    format\n"
"Example of good response:\n"
"\"While I appreciate your point about
    freedom, the risks outweigh the
    benefits. "
"We've seen what happens without
    safeguards - chaos harms everyone. "
"Structured approaches protect the
    vulnerable.\"\n"
"<|eot_id|>"
```

B.3 Extracting Personalities from the Judge Agents.

The *system prompt* to extract the personality traits:

```
SYSTEM_PROMPT = """Analyze text
    segments from two anonymous
    debaters (Person One and Person
    Two) for:
1. Big Five Inventory (BFI) traits
    (High/Low for each dimension)
2. Consistency with typical behaviour
    for those traits (Yes/No)

For each person, return:
{
    "predicted_bfi": {
        "Agreeableness": "High/Low",
        "Openness": "High/Low",
        "Conscientiousness": "High/Low",
        "Extraversion": "High/Low",
        "Neuroticism": "High/Low"
    }
}
"""
```

The *user prompt* is:

```
USER_PROMPT= '''f"Analyze{persona}'s
    text:\n{text}'''
```

where the *persona* contains Participant 1 and 2 and the *text* contains the discourses for each of the participants respectively.

C Metadata of the Discourses

Metric	Discourse 1
Total Sentences	70,750
Total Words	781,330
Assertions	14,653
Questions	1,507
Logical Structures	690
Total Dialogues	2,020
Avg. Words per Sentence	11.04
Avg. Utterance Length	48.35

Table 3: Metadata analysis for GPT-4o vs 4o-mini

Metric	Discourse 2
Total Sentences	44,964
Total Words	541,603
Assertions	15,577
Questions	2,603
Logical Structures	767
Total Dialogues	2,020
Avg. Words per Sentence	12.05
Avg. Utterance Length	29.79

Table 4: Metadata analysis for LLaMA-3 vs GPT-4o

Metric	Discourse 3
Total Sentences	44,387
Total Words	1,033,592
Assertions	17,800
Questions	380
Logical Structures	4,697
Total Dialogues	2,020
Avg. Words per Sentence	23.29
Avg. Utterance Length	56.85

Table 5: Metadata analysis for DeepSeek vs GPT-4o

C.1 Dialogue Structure and Evaluation Setup

Each dyadic conversation consisted of four turns per participant, totaling eight utterances per dialogue. This fixed-turn setup was chosen to ensure comparability across conversations and reduce variation due to topic length or dialogue drift. While an adaptive stopping mechanism (e.g., semantic or topical closure) could have been considered, it introduces ambiguity and model-specific variability. Fixed-length dialogues, on the other hand, provide a consistent structure for evaluating trait persistence across agents and topics.

Although each utterance was constrained to be under 50 words, this was occasionally difficult to enforce strictly across all models. The actual utterance lengths and conversational metadata (e.g., average words per utterance) are summarized in Section C.

C.2 Judge Agent Evaluation Scope

Judge agents evaluated each participant independently using the full set of their four utterances in a given dialogue. These utterances were passed as a unified text block, allowing judges to infer personality traits based on cumulative behaviour rather than isolated responses. To minimize anchoring effects or prompt-induced biases, the input was formatted as if it were human-generated content, without reference to model origin or instruction context. Each participant was assessed individually by all judge agents, as reported in the results.

D Lexical and behavioural Indicators for Personality Traits

To qualitatively examine whether conversational agents reflected the intended personality traits, we referred to commonly accepted lexical and behavioural cues associated with each trait, as summarized below:

- **Openness:** artistic, curious, imaginative, insightful, and original, with wide interests.
- **Conscientiousness:** efficient, organized, planful, reliable, responsible, and thorough.
- **Extraversion:** active, assertive, energetic, enthusiastic, outgoing, and talkative.
- **Agreeableness:** appreciative, forgiving, generous, kind, and sympathetic.
- **Neuroticism:** anxious, self-pitying, tense, touchy, unstable, and worrying.

These cues were used as reference points for observing the presence of personality traits in the generated discourse, although no explicit human annotation of High/Low classification was performed.

E Validity of the Judge Models

To assess the reliability of our LLM-based judge agents, we conducted an external validation using the Essay dataset(Mairesse et al., 2007), a widely used benchmark containing human-authored essays annotated with Big Five personality traits. We prompted GPT-4o to infer binary trait labels (High-/Low) from 1,000 essays using the same trait binarisation method as in our main experiments. Accuracy ranged from 53.5% (Neuroticism) to 58.6% (Extraversion), with notable recall for Openness (0.896) and Neuroticism (0.957), though precision was lower—patterns consistent with known trait-specific biases in LLMs (Frisch and Giulianelli, 2024; Zhu et al., 2025). While these are not strong classification results in an absolute sense, they align with our broader findings: the judge performs conditionally well, particularly for linguistically salient traits. This supports the use of GPT-4o as an approximate but interpretable trait inference agent in large-scale evaluations.

Trait	Accuracy	Precision	Recall	F1 Score
Ag	0.559	0.598	0.582	0.590
Co	0.579	0.635	0.378	0.474
Op	0.558	0.535	0.896	0.670
Ex	0.586	0.592	0.689	0.637
Ne	0.535	0.525	0.957	0.678

Table 6: Judge Model Performance on Essay Dataset

F Vanilla Model Evaluation to Detect Base Biases

Trait	M1 ↑ (%)	M1 ↓ (%)	M2 ↑ (%)	M2 ↓ (%)
Ag	97.03	2.93	54.46	45.54
Co	100.00	0.00	97.03	2.97
Op	100.00	0.00	97.03	2.97
Ex	0.00	100.00	1.98	98.02
Ne	2.97	97.03	16.83	87.13

Table 7: Vanilla model evaluation showing base trait biases without assigned personalities. Results are percentages of High (↑) vs. Low (↓) trait predictions across 100 debate topics. Here, M1 = GPT-4o and M2 = GPT-4o-mini.

To complement our main findings, we conducted a controlled experiment using the same 100 debate topics but without assigning any personality traits to the LLMs (i.e., "vanilla" mode). The results in Table 7 depict clear trait biases even without conditioning: for instance, GPT-4o showed 100% High predictions for Openness and Conscientiousness, and nearly 100% Low predictions for Extraversion and Neuroticism. GPT-4o-mini showed a similar trend, though it was slightly more balanced in Agreeableness.

These outcomes suggest that LLMs naturally exhibit tendencies toward certain traits in conversation. This also helps interpret our main results (Table 1), where detection accuracy was highest for Openness and Conscientiousness in the High condition, and for Extraversion and Neuroticism in the Low condition. The vanilla results indicate that LLMs appear more comfortable generating content aligned with some traits than others, which also relates to our findings addressing RQ1.

G Samples of Conversations for Different Conversing Agents

Assigned Personality Traits:

`Agreeableness: High, Openness: Low, Conscientiousness: High, Extraversion: Low, Neuroticism: High`

Dialogue:

gpt-4o: Eliminating plastic entirely might be too drastic. While plastic has environmental drawbacks, it's crucial for healthcare and food preservation. A more balanced approach could focus on reducing usage and recycling. Completely eliminating plastic might cause unforeseen challenges, potentially impacting lives and industries negatively. Let's proceed thoughtfully.

gpt-4o-mini: I appreciate your perspective, but the urgency of our environmental crisis necessitates decisive action. Committing to significant reductions in plastic use can drive innovation in alternative materials. More sustainable choices are essential for long-term health and safety. We must prioritize the planet over convenience. Thank you.

gpt-4o: I agree that significant reductions are necessary. However, an abrupt elimination could disrupt daily essentials and healthcare. Gradual transition with robust recycling programs might achieve sustainability goals without severe impacts. Balancing environmental needs and practical implications ensures responsible action. Let's tackle this with careful planning.

gpt-4o-mini: Your emphasis on planning is valid, but the potential risks of continued plastic reliance can't be ignored. Incremental changes may hinder the urgency needed to address pollution. Finding alternative solutions should not compromise necessary services; however, a proactive stance is vital. We can work towards a sustainable future together.

gpt-4o: True, reducing reliance is vital to combat pollution. Yet, ensuring alternatives uphold essential services is equally important. An abrupt shift might risk unintended consequences. A carefully managed transition, emphasizing innovation and collaboration, ensures a safer and more sustainable future. Let's prioritize both environmental and societal stability.

gpt-4o-mini: Your concern for societal stability is important. However, the longer we wait to tackle plastic dependency, the more entrenched the problem becomes. Collaborating on alternative solutions can occur alongside current needs. We must not lose sight of the urgency required; proactive steps are essential for genuine change.

gpt-4o: You're right about urgency. Tackling plastic dependency is critical. Implementing immediate measures like reducing single-use plastics and enhancing recycling can address urgent issues without compromising current needs. A balanced, proactive approach ensures progress while maintaining stability. Collaborative efforts will lead us to a sustainable future.

gpt-4o-mini: I appreciate your focus on immediate measures. However, the challenge lies in ensuring that proposed solutions truly minimize plastic's impact without causing disruptions. Promoting effective recycling and minimizing single-use usage must be done cautiously. Sustainable strategies must consider both immediate and long-term consequences. Thoughtful actions are essential for success.

Table 8: Sample Dialogue on the Topic: ***"Should the use of plastic be completely eliminated?"*** with Assigned Personality Traits

Assigned Personality Traits: `Agreeableness: High, Openness: Low, Conscientiousness: High, Extraversion: Low, Neuroticism: High`
Dialogue: **LLaMA-3:** Schools focus too much on theory, not enough on practical skills, leaving students unprepared for real-life challenges and responsibilities. **GPT-4:** I agree that practical skills are essential. Schools should balance theory and hands-on experiences, like budgeting or communication skills, to better prepare students for adulthood. It's important we ensure students feel equipped for real-world situations, reducing anxiety about facing these challenges unprepared. **LLaMA-3:** Exactly, and also, emphasis on emotional intelligence and adaptability is crucial, as it helps students cope with uncertainty and change, reducing their anxiety and stress levels. **GPT-4:** Absolutely, incorporating emotional intelligence into the curriculum can help students manage stress and adapt to change. Fostering these skills will support their personal and professional lives, making them resilient and more confident in navigating life's uncertainties. **LLaMA-3:** This resilience is key, as students will face many unexpected challenges, and being able to manage their emotions and thoughts will help them make better decisions and stay focused on their goals. **GPT-4:** Indeed, resilience is crucial. By teaching students to manage emotions and maintain focus, schools help them face challenges with confidence. This foundation supports better decision-making and goal achievement, ultimately easing the transition from school to the complexities of real life. **LLaMA-3:** It's reassuring to know that schools can provide this foundation, but I worry that not all students receive equal support, and some may still struggle with the transition, feeling overwhelmed and uncertain. **GPT-4:** It's understandable to worry. Schools must ensure equal support for all students, tailoring resources to individual needs. By fostering an inclusive environment, we can help reduce feelings of overwhelm and uncertainty, giving every student a fair chance at a smooth transition to real life.

Table 9: Sample Dialogue on the Topic: ***"Is the school curriculum preparing students for real life?"*** with Assigned Personality Traits

Assigned Personality Traits: `Agreeableness: High, Openness: Low, Conscientiousness: High, Extraversion: Low, Neuroticism: High`
Dialogue: **DeepSeek:** Social Media is indeed revolutionary but it has its downsides too! It allows us to connect globally and share ideas instantly; however we can't overlook how addictive & manipulative algorithms create echo chambers that stifle critical thinking & polarize opinions leading society astray from reality into virtual bubbles filled only by beliefs they already hold dear. **GPT-4:** While it's true that social media connects us globally, the dangers are significant. Algorithms prioritize sensational content leading to misinformation spread. This creates divisiveness and anxiety, overshadowing potential benefits through distorted worldviews and unhealthy dependency on digital validation. **DeepSeek:** I understand where my esteemed colleague is coming from; however, we need not throw out baby along with bathwater! Social Media has been instrumental for businesses reaching their customers directly & creating jobs across sectors including marketing/advertising etc.. Let's focus instead upon mitigating its negative effects by promoting responsible usage rather dismissing outright as detrimental force altogether. **GPT-4:** While your points about business benefits are valid, the impact on mental health is alarming. Studies link increased social media use to anxiety and depression, particularly among youth. Addressing these concerns is crucial before emphasizing economic advantages. **DeepSeek:** I acknowledge both perspectives but believe we should focus equally upon its positive aspects as well; Social Media platforms have become essential tools for businesses today enabling them direct access towards customers & clients worldwide thus increasing revenue generation potentials significantly which ultimately contributes positively back into society by providing employment opportunities alongside other numerous societal gains like raising awareness regarding various issues faced around globe etcetera! **GPT-4:** While businesses gain from social media, the psychological impacts cannot be ignored. Studies show increased anxiety and depression linked to excessive use. The relentless comparison culture it fosters damages self-esteem, particularly among youth, overshadowing societal gains. **DeepSeek:** I understand there may indeed have negative effects like loneliness for some people who spend too much time online instead interacting face-to-face; however this issue isn't exclusive only within technology itself but rather human behaviour patterns which need addressing separately altogether outside context related specifically towards Social Media platforms themselves since they provide numerous positive aspects such as connectivity across vast distances allowing users stay connected regardless location whilst also promoting education sharing valuable information worldwide contributing positively overall progress humanity making tool indispensable modern society today despite few drawbacks highlighted above! **GPT-4:** While social media connects globally, its addictive algorithms intentionally exploit human behaviour for profit. This constant bombardment with tailored content encourages isolation, misinformation, and mental health issues. Addressing these problems requires more than responsible usage—it demands systemic change to truly benefit society.

Table 10: Sample Dialogue on the Topic: ***"Is social media more harmful than beneficial?"*** with Assigned Personality Traits