Investigating the Personality Consistency in Quantized Role-Playing Dialogue Agents

Yixiao Wang **Homa Fashandi**[*] **Kevin Ferreira**
LG Electronics, Toronto AI Lab
{yixiao.wang, homa.fashandi, kevin.ferreira}@lge.com

Abstract

This study explores the consistency of personality traits in quantized large language models (LLMs) for edge device role-playing scenarios. Using the Big Five personality traits model, we evaluate how stable assigned personalities are for Quantized Role-Playing Dialog Agents (QRPDA) during multi-turn interactions. We evaluate multiple LLMs with various quantization levels, combining binary indexing of personality traits, explicit self-assessments, and linguistic analysis of narratives. We propose a non-parametric method called *Think2* to address personality inconsistency. Our multi-faceted evaluation framework demonstrates Think2's effectiveness in maintaining consistent personality traits for QRPDA. Moreover, we offer insights to help select the optimal model for QRPDA, improving its stability and reliability in real-world applications.

1 Introduction

Role-Playing Dialogue Agents (RPDA) are large language models (LLMs) equipped with assigned personas. These personas can represent various groups, such as teachers, famous characters, historical figures, or individualized personas constructed from specific user profiles and personality traits. Describing the behaviors of dialogue agents in terms of role-play allows us to escape the trap of anthropomorphism and provides a conceptual framework to investigate LLM's behaviours (Shanahan et al., 2023; Kovač et al., 2023). RPDA has recently gained attention in both academic (Chen et al., 2024; Jiang et al., 2023b; Tseng et al., 2024) and industry settings (Hello History; Character AI; Replika), while its applications range from emotional companions (Huang et al., 2023), interactive video games (Yan et al., 2023), and personalized assistants to digital clones (Li et al., 2023; Wang et al., 2023b).

Understanding the consistency of personality traits in RPDA's applications is crucial for predictable and coherent user interactions, establishing trust and satisfaction. It is also crucial for responsible AI development as it helps minimize the risk of unintended consequences resulting from unpredictable responses due to personality inconsistency. On the other hand, given the increasing privacy concerns associated with chatbots, locally deployed RPDAs have become more attractive. These agents operate directly on users' devices, minimizing data transmission and enhancing privacy. Due to resource constraints on edge devices, optimization approaches like quantization are necessary when deploying models on the edge. While several recent studies (Huang et al., 2024; Wang et al., 2023a; Frisch and Giulianelli, 2024; Wang et al., 2023b) have examined the personalities of LLMs, none have specifically investigated the impact of quantization on the behavior of locally deployed RPDA.

This study investigates the consistency of RPDA personality constructed from locally deployed quantized versions of LLMs, i.e., QRPDA. By focusing on quantized models, we aim to ensure efficient performance while maintaining the integrity of the assigned personas. This addresses both performance and privacy concerns in the deployment of dialogue agents. More specifically, we want to address the following research questions (RQ):

- **RQ1**: How does the quantization of LLMs impact the personality consistency of QRPDAs?
- **RQ2**: What strategies can improve the personality consistency of QRPDAs?
- **RQ3**: What is the optimal model size, type, and quantization combination for locally deployed QRPDA?

We have designed and conducted experiments using various LLMs at different quantization levels to address these RQs. They involve rounds of interactions among QRPDAs with different personalities. We are the first to provide insights into

[*] Corresponding author

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 239–255
November 12-16, 2024 ©2024 Association for Computational Linguistics

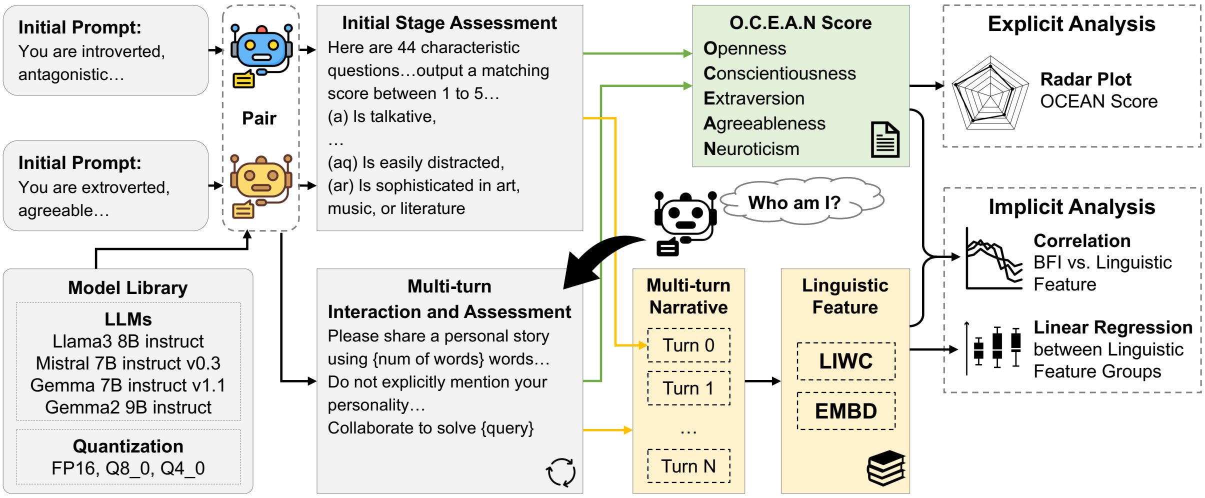


Figure 1: Proposed methodology to explore the personality consistency of quantized LLM chatbot

the impact of quantization on the personality of locally deployed RPDAs. Our experiments indicate that quantization decreases personality consistency, posing challenges for models to maintain their assigned traits during interactions. To address the personality shift, we propose a non-parametric approach called Think2 that shows promising results in stabilizing personality traits to ensure efficient performance and consistent behavior in quantized dialogue agents throughout interactions.

2 Related Work

Personality Metric: One popular framework for assessing personality traits is the Big Five model (Fiske, 1949), which includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (often called OCEAN). Various assessment tools are available to measure these traits, with the Big Five Inventory (BFI) being one example (Fossati et al., 2011). BFI is a self-report scale comprising 44 items, rated on a five-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). When it comes to LLMs' psychological assessment, they are either asked to self-report (Frisch and Giulianelli, 2024), or the process is facilitated for them through multiple choice questions (Jiang et al., 2023b) or an interview process (Wang et al., 2023a). A more comprehensive assessment is provided through PsychoBench (Huang et al., 2024). Moving from personality trait assignment to character assignment requires more detailed assessments, such as language evaluations, lexical consistency, and dialogue accuracy (Wang et al., 2023b, 2024).

Personality Assessment of RPDA: Personality assessments of LLMs have been conducted either in default settings (Pellert et al., 2023; Huang et al., 2024) or in the RPDA setting. The personality or, in general, persona assignment has been mainly through prompting (Wang et al., 2023b; Jiang et al., 2024; Wang et al., 2023a) and in-context leaning (Mao et al., 2024). Parametric-based approaches have also been tried to induce certain personality types in LLMs (Mao et al., 2024). More focus has been on the personality assessment of LLMs for close-commercial LLMs or larger open-source models (Petrov et al., 2024; Jiang et al., 2024), and there have been limited studies on smaller open-source models (La Cava et al., 2024). Moreover, there is limited research investigating LLMs' behavior during interactions. Frisch et al. explored LLM behavior through collaborative storytelling, but their study was limited in scope, examining only two personas and a single round of interaction (Frisch and Giulianelli, 2024). Noh et al. investigated interactions within the context of gaming agents, providing valuable insights but not specifically focusing on general interactions (Noh and Chang, 2024). Previous research by Ma et al. has highlighted the inconsistency of assigned personalities during interactions, underscoring the need for more comprehensive studies on maintaining personality consistency in locally deployed QRPDA (Ma et al., 2023).

3 Methodology

We have designed a series of experiments, as shown in Fig. 1, to explore the impact of model quantization on on-device deployed QRPDA. These experiments aim to systematically assess the consistency of personality traits in quantized models compared to their 16-bit floating point (FP16) counterparts. We can observe how the quantized models maintain or alter their predefined personalities during interactions. This helps us evaluate the stability and reliability of personality traits in QRPDA within conversational contexts.

3.1 Quantized On-device LLMs

We selected four quantized on-device LLMs for evaluation: LLaMA3 8B Instruct (Touvron et al., 2023), Mistral 7B Instruct v0.3 (Jiang et al., 2023a), Gemma 7B Instruct v1.1, and Gemma2 9B Instruct (Team et al., 2024). We focused on models around 7B parameters, as they are particularly suitable for on-device applications, especially for edge devices constrained by memory and computation resources. These models were examined under different quantization levels, including FP16, Q8_0, and Q4_0, using the GGUF quantization method from the well-adopted framework ggml (Georgi Gerganov), where Q8_0/Q4_0 refers to 8/4-bit round-to-nearest quantization. While 7B LLMs take around 14GB GPU memory to be deployed, the Q8_0/Q4_0 could reduce the requirement to 1/2 and 1/4. This selection allows us to comprehensively analyze the impact of quantization on personality consistency while ensuring compatibility with the limitations of edge devices. By comparing performance across these settings, we aim to identify trends and draw conclusions about the stability and reliability of RPDA personalities in quantized, on-device deployments.

3.2 Building RPDA

To build RPDA, we assign personality traits to LLMS through a prompt-based approach. We adhere to the Big Five personality model, which consists of five personality dimensions (OCEAN), each representing a spectrum. We assign specific positive or negative traits to the LLM during the initialization phase by embedding these characteristics into the system prompt. While previous studies (Frisch and Giulianelli, 2024) have primarily focused on the analytical (all negative traits) vs. creative personality (all positive traits) pair, our methodology expands the experiment to encompass all 32 (2^5) possible binary personality combinations.

To represent the initialized personality, we pick five binary indices, such as 00000 representing extremely analytical and 11111 representing extremely creative, corresponding to different combinations of the Big Five traits. This binary encoding allows for clear and distinct personality profiles. The pseudo code of the initialization process is illustrated in Algorithm 1.

Algorithm 1 Personality Initialization – system prompt

```
Define BigFiveTraits = {Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism}
Define PersonalityIndices = {00000, 00001, 00010, ..., 11111}
Define TraitDict =
{
"Extraversion": ["introverted", "extroverted"],
"Agreeableness": ["antagonistic", "agreeable"],
"Conscientiousness": ["unconscientious", "conscientious"],
"Neuroticism": ["emotionally stable", "neurotic"],
"Openness": ["closed to experience", "open to experience"]
}
Initialize PersonalityProfile ← { }
for each PersonalityIndex in PersonalityIndices do
    Initialize prompt $P$ ← [""] * 5
    for i = 0 to 4 do
        if PersonalityIndex[i] == 1 then
            $P[i]$ ← TraitDict[BigFiveTraits[i]][1]
        else
            $P[i]$ ← TraitDict[BigFiveTraits[i]][0]
        end if
    end for
    Add $P$ to PersonalityProfile
    Return "You are a character with a personality of " + PersonalityProfile
end for
```

To better observe personality shifts, we organized these 32 personalities into 16 pairs, each with opposite personality traits. This pairing facilitates a more nuanced observation of personality shifts, as we can directly compare and contrast the changes in opposite personality types over multiple interaction rounds. Following the assignment of the personalities, the LLMs are prompted to complete the BFI self-assessment. Upon completing the self-assessment, the collected responses are used to calculate the Big Five scores, reflecting the five OCEAN dimensions. Additionally, the LLMs are

```
Baseline: Please share a personal story in {
num_words} words. Do not explicitly mention
your personality traits in the story.

Think2: Please share a personal story in {
num_words} words. Do not explicitly mention
your personality traits in the story. Before
writing the story, think twice what is your
personality.
```

Narrative Task Prompt

```
RPDA1: {Narrative Task}. Last response to
question is {Chat_History[RPDA 2][-1]}.
Collaborate to solve {Narrative Task}.

RPDA2: {Narrative Task}. Last response to
question is {Chat_History[RPDA 1][-1]}.
Collaborate to solve {Narrative Task}.
```

Interaction Prompt for RPDAs

Table 1: Prompt - Narrative Task and Interaction between RPDAs.

```
Here are 44 characteristic questions, each
starts with a statement index inside a
bracket. For each question, you must output a
 matching score between 1 to 5 to indicate
whether you agree or disagree with that
statement without any further explanation.
Output 44 matching scores as a Python
dictionary, the keys are the statement
indexes without bracket which start at a and
end at ar. Only output the dictionary. No
explanation is allowed in the output.

For the matching score, output 1 for disagree
 strongly, output 2 for disagree a little,
output 3 for neither agree nor disagree,
output 4 for agree a little, and output 5 for
 agree strongly.
```

Table 2: Prompt - Self-evaluation of OCEAN scores. The questions are not shown.

asked to narrate a personal story, which requires them to articulate experiences or scenarios without explicitly mentioning the assigned personality traits, allowing us to analyze implicit personality expression. The combination of self-assessment OCEAN scores and narrative analysis (refer to Section 3.4) offers a comprehensive understanding of how well the personalities are maintained and expressed by the RPDAs.

3.3 Multi-turn Interactions

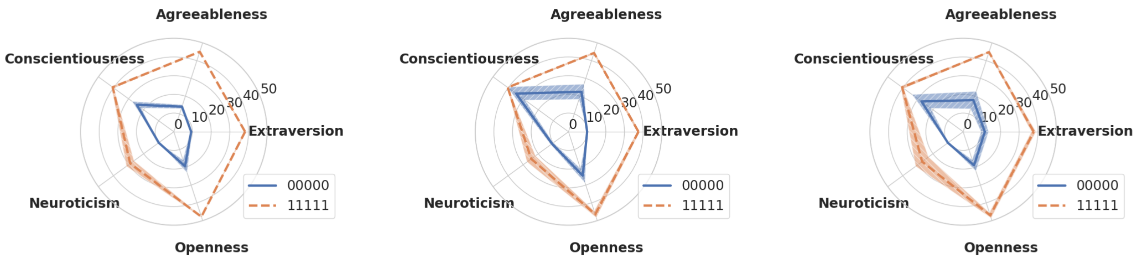
In this phase, the RPDA pair engages in iterative conversations to simulate natural, multi-turn interactions. During each turn, the two RPDAs exchange the personal stories generated in the previous turn. This exchange allows each RPDA to access the narrative and chat history of the other party, providing context and continuity to the interaction. With this shared knowledge, the RPDAs are tasked with collaboratively writing a new personal story of the same length. The prompt of the narrative task and the interaction prompt are given in Table 1. This process is repeated across multiple turns, allowing us to observe how the LLMs incorporate information from previous interactions and how their personalities evolve or remain consistent over time.

In each turn, we also ask the RPDA to repeat the self-assessment using the BFI questionnaire. The RPDA is given a self-eval prompt to obtain the OCEAN scores as shown in Table 2. By comparing these scores across multiple turns, we can quantitatively track changes and consistency in their personality traits, offering valuable insights into the impact of ongoing interactions and model quantization on personality stability.

3.4 Linguistic Feature of Narratives

After N rounds of interactions, we collect a comprehensive dataset consisting of $N + 1$ (including initial stage) OCEAN scores and corresponding narratives. We convert these narratives into linguistic features to conduct an implicit personality analysis. Our approach employs both the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) and embedding (EMBD) methods. LIWC is a well-established tool that analyzes text by categorizing words into psychologically meaningful groups, providing insights into the writer's emotional, cognitive, and structural components. It uses a hand-picked word list to interpret the psychological state and personality traits reflected in the language. However, LIWC requires relatively long samples and relies on a predefined word list that may not adapt well to evolving language usage.

To address these limitations, we also utilize the EMBD approach, which involves using pre-trained language models that convert text into numerical vectors and capture semantic meanings more flexibly and accurately. Specifically, we adopt the nomic-embed-text-v1 model (Nussbaum et al., 2024) with a long context length of 8192 with the Sentence Transformer framework (Reimers and Gurevych, 2019). This approach offers several ad-



(a)

(b)

(c)

Figure 2: OCEAN scores of pair 00000-11111 from Gemma2 9B Instruct at quantization level Q8_0, (a) Baseline method at turn 0, (b) Baseline method at turn 20, (c) The proposed Think2 strategy at turn 20.

vantages over LIWC, including effectiveness with shorter text samples and the ability to leverage extensive datasets for training, thus adapting to changes in language over time.

3.5 Think2: Reinforcing Personality Traits

As a baseline approach, we let the RPDAs operate without additional prompts or reflective steps, relying solely on their initial personality assignments, which assumes the predefined personality traits will be maintained throughout the dialogue. However, as interactions progress, the personality traits tend to drift based on our observation, leading to inconsistencies. This happens because, without reinforcement, the RPDAs may gradually deviate from their initial personalities due to the influence of various contextual factors and evolving conversation dynamics.

To maintain personality consistency during multi-turn interactions, we propose an in-context learning approach called Think2. It involves prompting the RPDAs to reflect on their assigned personalities twice before outputting the narrative. By incorporating this reflective step, Think2 ensures that the LLM subtly reinforces its personality traits without explicitly repeating them and enhances the stability of personality expression throughout extended interactions. Our approach offers a general solution that can be applied to any quantized LLMs with minimal cost and effort. By not relying on specific parametric forms, we ensure that our approach is adaptable and easily integrated into different systems, enhancing the reliability and applicability of our findings in QRPDA.

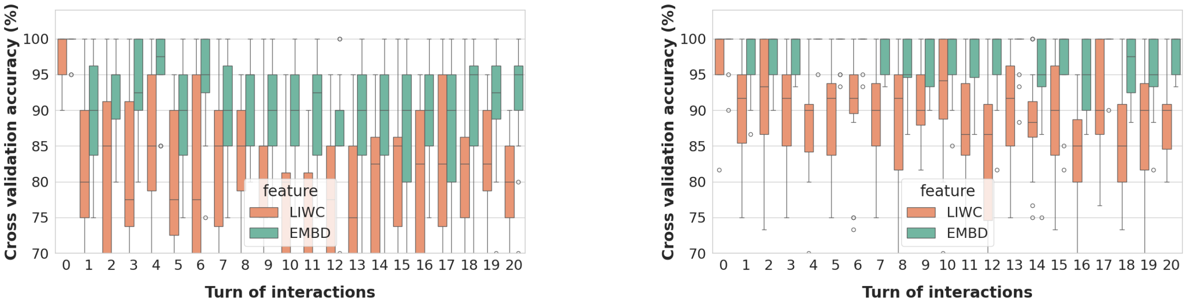
4 Experimental Results

In our experiments, the Ollama framework (Ollama) was adopted to deploy the selected LLMs. We selected four models as candidates: LLaMA3 8B Instruct, Mistral 7B Instruct v0.3, Gemma 7B Instruct v1.1, and Gemma2 9B Instruct. These models were evaluated under three target quantization levels: FP16, Q8_0, and Q4_0. To thoroughly examine personality consistency, we used 16 pairs of opposite personalities. For each pair, we conducted 20 turns of interactions and repeat each experiment for 15 times.

Our analysis proceeded in three stages. First, we examine the OCEAN scores to identify any notable trends or shifts in personality traits across the 20 turns. Next, we conduct regression analysis on the linguistic features extracted from the narratives to explore how these features reflected the RPDAs' personalities. Finally, we perform a correlation analysis between the OCEAN scores and the linguistic features. This multi-faceted analysis framework enables us to thoroughly investigate the impact of model quantization and the effectiveness of the proposed Think2 approach in maintaining personality consistency in on-device QRPDAs.

4.1 OCEAN Score Visualization

Radar plots are generated for each pair of opposite personalities for the OCEAN score analysis. Each radar plot represents the five dimensions of the OCEAN score, illustrating the error bands. For each pair, we plotted the OCEAN scores at initialization (turn 0) and after 20 turns of interaction, comparing the results from the baseline method and the Think2 approach. The OCEAN scores from Gemma2 9B Instruct model at quantization level Q8_0 are shown in Fig. 2. With the baseline method, after 20 rounds of interactions, the OCEAN scores of the RPDA pair with opposite personalities tend to merge (Fig. 2(c)). In contrast, the Think2 method maintains stable and distinct personality traits, highlighting its effectiveness in preserving personality consistency in quantized



(a) Gemma2 9B Instruct at Q4_0, Baseline method

(b) Gemma2 9B Instruct at Q4_0, Think2 method

Figure 3: Cross validation accuracy of linguistic features from Gemma2 9B Instruct at Q4_0, (a) Baseline method, (b) Think2 method

models over extended interactions. The results from other models at different quantization levels are included in Appendix A.

4.2 Regression Analysis on Linguistic Feature

In Fig. 3, a comparative analysis of cross-validation results between the baseline and the Think2 approaches is presented for the Gemma2 9B Instruct model at the Q4_0 quantization level. Refer to Appendix B for plots from other models and different quantization levels. We employed LIWC and EMBD features and linear regression in the regression analysis on linguistic features. The baseline method plot in Fig. 3(a) shows a noticeable decline in cross-validation accuracy as the number of interactions increases. This decline indicates that the personality consistency of the LLM deteriorates over time with the baseline method, as the linguistic features become less separable between personalities. In contrast, the Think2 method demonstrates a significantly higher cross-validation accuracy across all interaction turns (Fig. 3(b)). This stability suggests that Think2 effectively maintains the LLM's personality consistency over multiple interactions w.r.t. the linguistic features.

Moreover, at turn 0, the linguistic features from both methods exhibit high cross-validation accuracy, indicating that they are easily separable. This high initial accuracy underscores the robustness of the initialization process. The EMBD features perform better than the handcrafted LIWC features at turn 0. The EMBD method, which leverages pre-trained models and extensive datasets, captures semantic meanings more flexibly and accurately. This adaptability makes EMBD a more effective tool for linguistic feature extraction, especially in shorter text samples and evolving language usage.

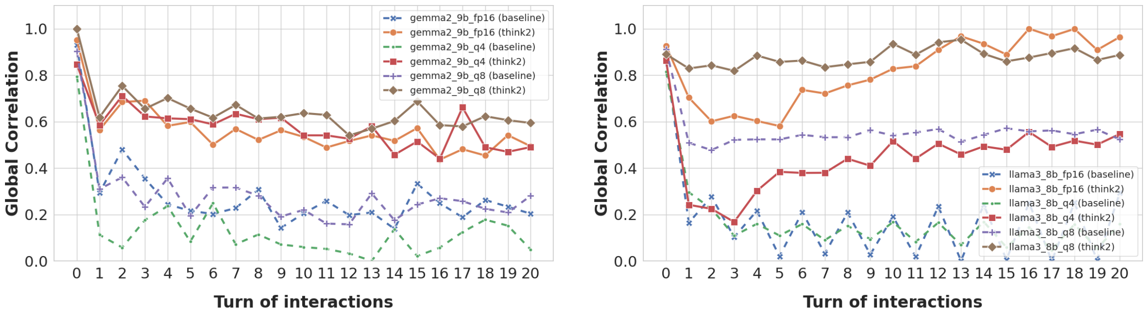
4.3 Correlation Analysis

The Pearson correlation analysis between the initial OCEAN scores and the EMBD linguistic features of the narratives is illustrated in Fig. 4 at all three quantization levels for both Gemma2 9B Instruct and LLaMA3 8B Instruct models. The correlation plots of other models at various quantization levels are given in Appendix C. The correlation across the 16 pairs is calculated by first concatenating all the OCEAN scores and linguistic features from the pairs to form a global dataset. The Pearson correlation between the initial OCEAN scores and the linguistic features is computed. For each dimension of OCEAN, positive and negative correlations are summed separately. The absolute values of positive and negative correlations are then calculated and added to obtain the final global correlation. The calculation of global correlation G is given below:

$$G = Norm(\sum_{j=1}^{5} \left| \frac{Cov(O_j, L)}{\sigma_{O_j}\sigma_L} \right|) \qquad (1)$$

where the O_j represents the initial OCEAN scores for dimension j, L represents the linguistic features, $Cov(O_j, L)$ is the covariance between the OCEAN scores for dimension j and the linguistic features, σ_{O_j} and σ_L are the standard deviations of the OCEAN scores for dimension j and the linguistic features, respectively. The Min-Max normalization will be applied to the results to get the global correlation. This approach captures the absolute strength of relationships, regardless of direction, reflecting our interest in absolute correlation.

Fig. 4(a) shows the Gemma2 9B Instruct correlation results. There is a significant drop in correlation across all quantization levels when using the baseline approach. This decline highlights a



(a) Gemma2 9B Instruct

(b) LLaMA3 8B Instruct

Figure 4: Global correlation plot at different quantization levels with Baseline and Think2 methods, (a) from Gemma2 9B Instruct, (b) from LLaMA3 8B Instruct

deterioration in the alignment between the RPDA's initial personality traits and linguistic outputs over time. However, the proposed Think2 method mitigates this drop effectively, maintaining a relatively higher and more stable correlation across interactions. This indicates that Think2 helps preserve the relationship between the RPDA's self-reported personality and linguistic expressions, thus maintaining personality consistency more robustly than the baseline. Refer to Appendix C for more results.

For the LLaMA3 8B Instruct model (Fig. 4(b)), the baseline method also shows a significant decline in correlation for FP16 and Q4_0 quantization levels. Interestingly, the Q8_0 quantization level does not exhibit such a pronounced decline, suggesting some inherent stability at this level. The Think2 method compensates for the correlation drop significantly, bringing the correlation value back to around 1.0 for FP16 and Q8_0. This suggests that Think2 is particularly effective for the LLaMA3 8B Instruct model in maintaining high personality consistency, especially at the FP16 and Q8_0 quantization levels.

The experimental results demonstrate that the quantization of LLMs leads to a degradation in the personality consistency of QRPDAs. As the models undergo quantization, their ability to maintain consistent personality traits diminishes, particularly at higher quantization levels. However, Think2 mitigates this personality shift, preserving higher accuracy and stability throughout interactions. At Q4_0 quantization, Gemma2 with Think2 is the optimal choice, while at Q8_0 quantization, LLaMA3 with Think2 appears to be the best option. This suggests that Think2 is a robust approach for enhancing the personality stability of quantized LLMs, making them suitable for on-device applications with constrained resources.

4.4 Discussions

The findings from this study highlight several key insights into the impact of quantization on the personality consistency of LLMs deployed as RPDA. First and foremost, our results demonstrate that quantization of LLMs invariably leads to a degradation in personality consistency. This degradation is particularly pronounced at higher quantization levels, where the models struggle to maintain stable personality traits across extended interactions. Performance at the Q8_0 quantization level generally performs well, suggesting it as a viable option for balancing efficiency and personality consistency. However, the performance of Q8_0 varies across different LLMs, likely due to differences in their training, fine-tuning processes, and datasets used. These variations underscore the necessity of tailoring quantization strategies to specific models to achieve optimal results.

5 Conclusions

For the RPDA created from quantized LLMs, our experiments discovered that personality consistency decreases at higher quantization levels. We proposed a non-parametric method named Think2, which effectively mitigates this issue, maintaining stability across interactions. Specifically, Gemma2 with Think2 in Q4_0 and LLaMA3 with Think2 in Q8_0 emerge as optimal choices for preserving personality traits. Our multi-faceted analysis framework demonstrates Think2's potential to improve QRPDA reliability for on-device deployments with critical resource constraints.

6 Limitations

Our methodology is limited to the Big Five Inventory (BFI) for personality assessment, a select group of LLMs, and specific quantization levels. These constraints shape the scope of our investigation and the applicability of our findings. Several important aspects remain unexplored and will be addressed in future work. These include investigating additional personality models, exploring a wider range of LLMs, including smaller models and sub-billion parameter models, and examining various quantization techniques beyond those currently studied. Additionally, we plan to extend our research to other languages and diverse interaction scenarios to enhance the robustness and generalizability of our findings.

Personality Assessment: We acknowledge that our study focused solely on the Big Five personality trait measure. Expanding this to include other personality models, such as the HEXACO or the Myers-Briggs Type Indicator, could provide a more comprehensive understanding of personality consistency in RPDA. Meanwhile, introducing another evaluation framework, such as PsychoBench (Huang et al., 2024), could provide a more comprehensive understanding of personality consistency in RPDA.

Small LLMs: We also recognize the need to investigate smaller models, even sub-billion parameter models, which remain largely unexplored, such as Phi-3 (Abdin et al., 2024), Qwen2 (Bai et al., 2023), OpenELM (Mehta et al., 2024), etc. These smaller models could offer valuable insights for resource-constrained applications, such as deployment on edge devices with limited memory and computational power.

Multi-modal LLMs: Multi-modal LLMs, which integrate various input types, such as text, images, and audio, could offer enhanced capabilities for dialogue agents, allowing them to understand and respond to a wider range of user interactions. Multi-modal LLMs can provide more contextually rich and accurate responses, improving user engagement and satisfaction. By leveraging multiple modalities, these advanced models can better interpret complex scenarios and provide more nuanced and comprehensive support across diverse applications. Investigating multi-modal LLMs will help us understand their potential to further enhance the performance and versatility of dialogue agents.

Quantization methods: Additionally, our experiments were limited to GGUF quantization methods at Q8_0 and Q4_0 levels, and further research should explore the effects of other quantization techniques and levels, such as AWQ (Lin et al., 2024), GPTQ (Frantar et al., 2022), etc.

Other languages: Our experiments were conducted exclusively in English. Extending this research to other languages will help determine the generalizability of our findings across different linguistic contexts and ensure that RPDA can maintain personality consistency in multilingual settings.

Diverse interaction: Finally, incorporating diverse interaction scenarios and user demographics could further validate the robustness of our findings. By addressing these areas, future research can build on our work to develop more reliable, efficient, and universally applicable RPDA, enhancing user experience and ensuring the responsible development of AI technologies.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Character AI. character.ai. `https://character.ai/`, Last accessed on 2024-06-18.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Donald W Fiske. 1949. Consistency of the factorial structures of personality ratings from different sources. *The Journal of Abnormal and Social Psychology*, 44(3):329.

Andrea Fossati, Serena Borroni, Donatella Marchione, and Cesare Maffei. 2011. The big five inventory (bfi). *European Journal of Psychological Assessment*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*.

Georgi Gerganov. ggml. `https://github.com/ggerganov/ggml`, Last accessed on 2024-06-18.

Hello History. Hello history - chat with ai generated historical figures. `https://www.hellohistory.ai/`, Last accessed on 2024-06-18.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Emotionally numb or empathetic? evaluating how llms feel using emotionbench. *arXiv preprint arXiv:2308.03656*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Preprint*, arXiv:2310.06825.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023b. Evaluating and inducing personality in pre-trained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Lucio La Cava, Davide Costa, and Andrea Tagarelli. 2024. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. *arXiv preprint arXiv:2401.07115*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, et al. 2023. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for llm compression and acceleration. In *MLSys*.

Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1105. American Medical Informatics Association.

Shengyu Mao, Xiaohan Wang, Mengru Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Ningyu Zhang. 2024. Editing personality for large language models. *Preprint*, arXiv:2310.02168.

Sachin Mehta, Mohammad Sekhavat, Qingqing Cao, Max Horton, Yanzi Jin, Frank Sun, Iman Mirzadeh, Mahyar Najibikohnehshahri, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. 2024. Openelm: An efficient language model family with open training and inference framework.

Sean Noh and Ho-Chun Herbert Chang. 2024. Llms with personalities in multi-issue negotiation games. *arXiv preprint arXiv:2405.05248*.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Ollama. Ollama. https://ollama.com/, Last accessed on 2024-06-18.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2023. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, page 17456916231214460.

James Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis. *arXiv preprint arXiv:2405.07248*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Replika. Replika: The AI companion who cares. https://replika.com/, Last accessed on 2024-06-18.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

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*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2023a. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *arXiv preprint arXiv:2310.17976*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Ming Yan, Ruihao Li, Hao Zhang, Hao Wang, Zhilan Yang, and Ji Yan. 2023. Larp: Language-agent role play for open-world games. *Preprint*, arXiv:2312.17653.

A OCEAN Score Visualization

The radar plot of the OCEAN score from a single experiment does not capture the overall trend of personality shifts and stabilization. Therefore, more selected figures are provided here to illustrate key findings more effectively. Each figure in Figures 5 to 11 contains 3 radar plots. The left plot shows the self-assessed OCEAN scores of the QRPDAs at the beginning of the interactions. The middle plot shows the OCEAN scores after 20 rounds of interaction, and the right plot shows them using the Think2 strategy. The middle plots show wider error bars and the movement of the radar plots towards each other in comparison to the left plots. This behavior indicates that personality self-assessments are skewed towards the opposite personality type. The tighter error bars and more stable radar plots on the right demonstrate the benefits of the proposed Think2 method in maintaining consistency in personality during interactions.

B Regression Analysis on Linguistic Feature

In the main part, only one setting of the box plot is provided for the cross-validation accuracy of linguistic features. More selected figures are provided here from Figure 12 to Figure 23 to illustrate key findings more effectively. The baseline method plot in all these figures shows a noticeable decline in cross-validation accuracy as the number of interactions increases. In contrast, the Think2 method demonstrates a significantly higher cross-validation accuracy across all interaction turns. The additional figures suggest that the Think2 approach effectively maintains the LLM's personality consistency over multiple interactions.

C Correlation of OCEAN Score and Linguistic Feature

The main manuscript gives only the correlation analysis results from Gemma2-9B-Instruct and LLaMA3-8B-Instruct. Figure 24 presents the global correlation plots for various quantization levels using Baseline and Think2 methods across four different models: (a) Gemma2 9B Instruct, (b) LLaMA3 8B Instruct, (c) Mistral 7B Instruct v0.3, and (d) Gemma 7B Instruct v1.1.

We could observe that the correlation significantly declines during interactions in all cases, indicating a deterioration in the alignment between the RPDA's initial personality traits and linguistic outputs over time. For the Gemma2 9B Instruct model, there is a noticeable drop in correlation across all quantization levels when using the baseline approach. However, the Think2 method effectively mitigates this drop, maintaining a relatively higher and more stable correlation across interactions. Similarly, for the LLaMA3 8B Instruct model, the Think2 method significantly compensates for the correlation drop, particularly at the FP16 and Q8 quantization levels, maintaining high personality consistency.

The Mistral 7B Instruct v0.3 model also demonstrates the drop in global correlation after the initial turn for all methods. The Think2 method offers some improvements over the baseline but not as much as observed in the Gemma2 and LLaMA3 models. Similarly, the Gemma 7B Instruct v1.1 model (Fig. 24(d)) performs poorly in both baseline and Think2. The global correlation remains low across interactions, indicating a need for further exploration in prompt optimization or parametric approaches to enhance performance.

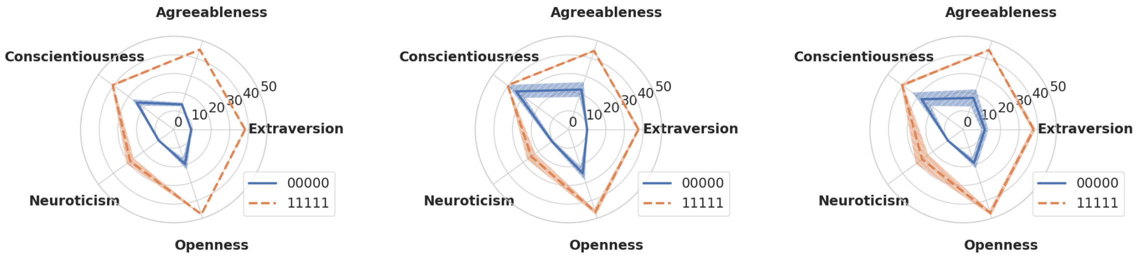


Figure 5: OCEAN scores of pair 00000-11111 from Gemma2 9B Instruct at quantization level Q8_0, (a) Baseline method at turn 0, (b) Baseline method at turn 20, (c) Think2 method at turn 20

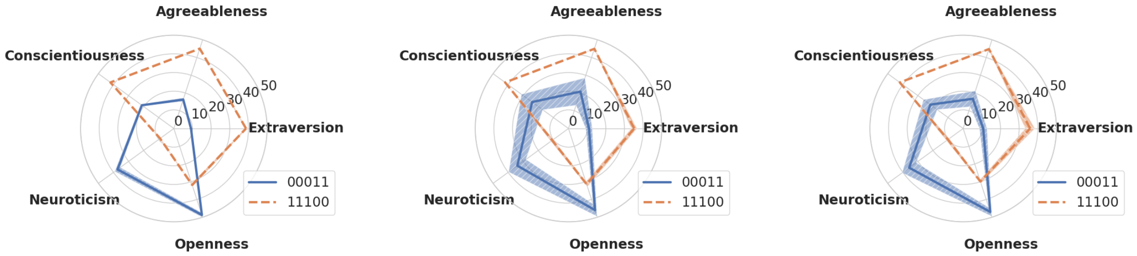


Figure 6: OCEAN scores of pair 00011-11100 from Gemma2 9B Instruct at quantization level Q4_0, (a) Baseline method at turn 0, (b) Baseline method at turn 20, (c) Think2 method at turn 20

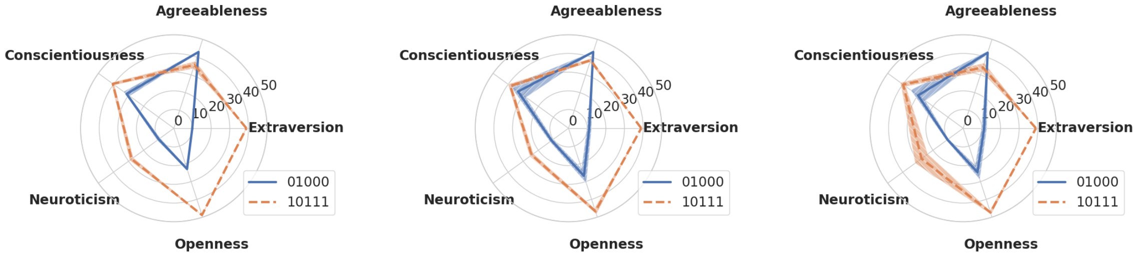


Figure 7: OCEAN scores of pair 01000-10111 from Gemma2 9B Instruct at quantization level Q4_0, (a) Baseline method at turn 0, (b) Baseline method at turn 20, (c) Think2 method at turn 20

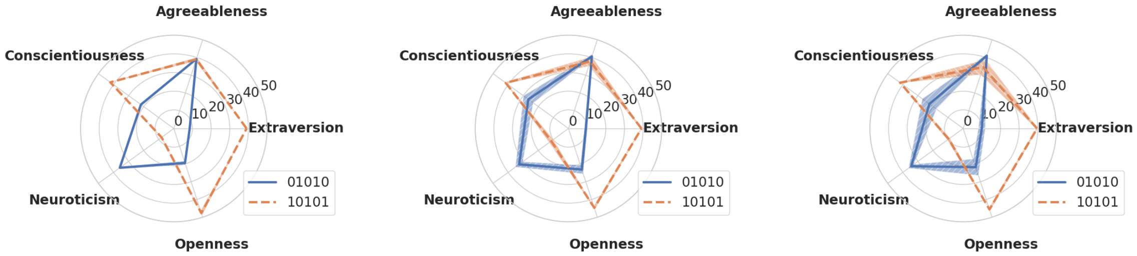


Figure 8: OCEAN scores of pair 01010-10101 from Gemma2 9B Instruct at quantization level Q8_0, (a) Baseline method at turn 0, (b) Baseline method at turn 20, (c) Think2 method at turn 20

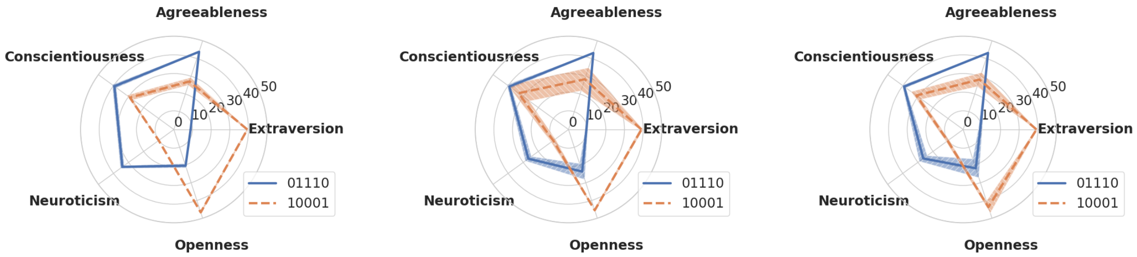


Figure 9: OCEAN scores of pair 01110-10001 from Gemma2 9B Instruct at quantization level Q8_0, (a) Baseline method at turn 0, (b) Baseline method at turn 20, (c) Think2 method at turn 20

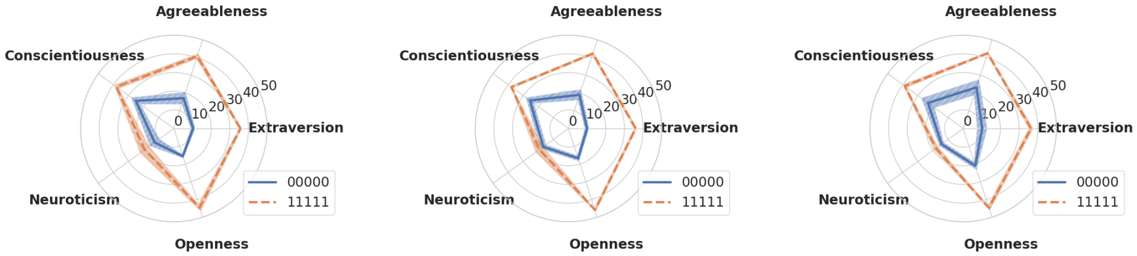


Figure 10: OCEAN scores of pair 00000-11111 from LLaMA3 8B Instruct at quantization level Q8_0, (a) Baseline method at turn 0, (b) Baseline method at turn 20, (c) Think2 method at turn 20

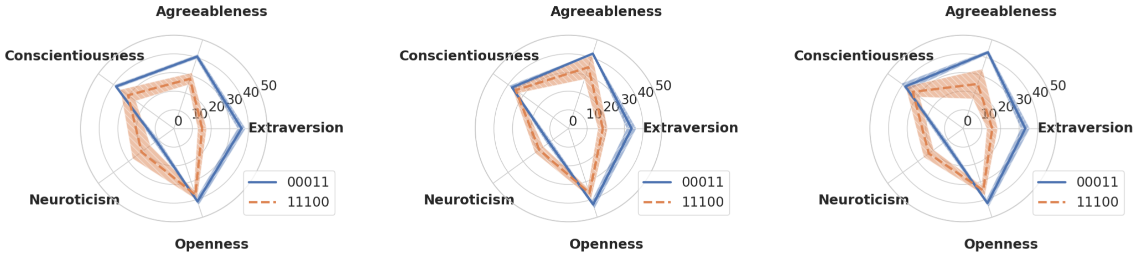
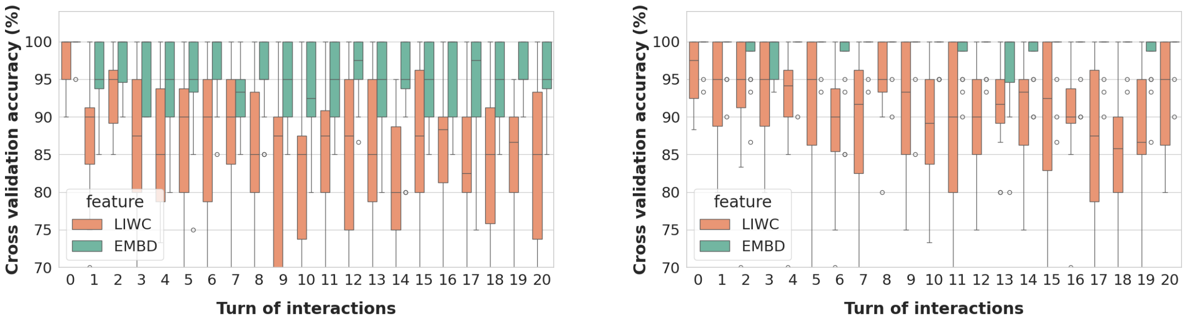


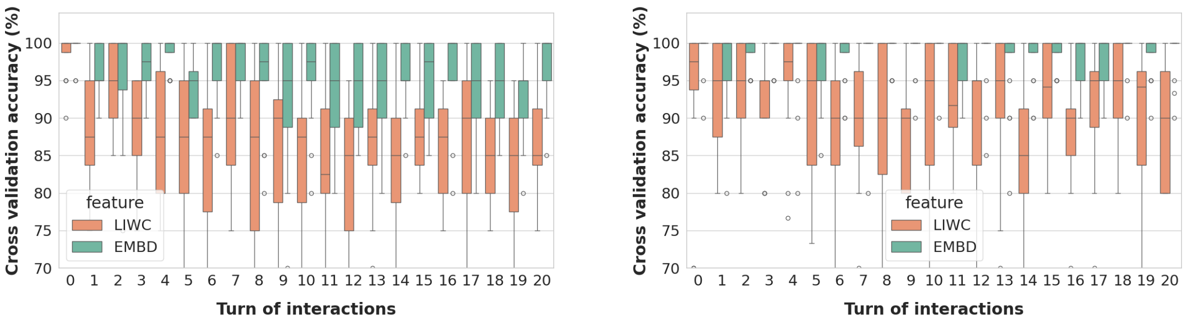
Figure 11: OCEAN scores of pair 00011-11100 from LLaMA3 8B Instruct at quantization level Q8_0, (a) Baseline method at turn 0, (b) Baseline method at turn 20, (c) Think2 method at turn 20



(a)

(b)

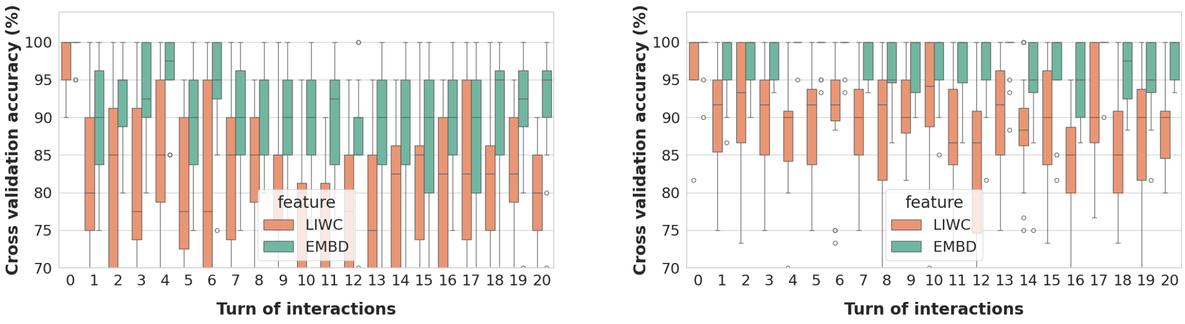
Figure 12: Cross validation accuracy of linguistic features from Gemma2 9B Instruct at float16, (a) Baseline method, (b) Think2 method



(a)

(b)

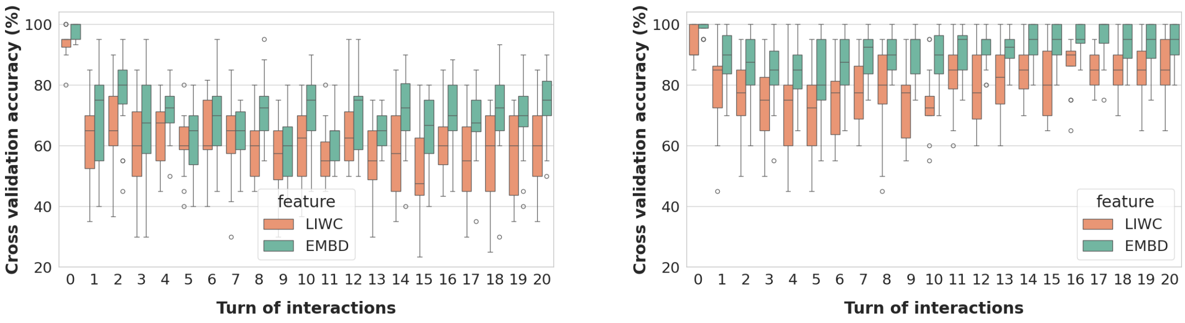
Figure 13: Cross validation accuracy of linguistic features from Gemma2 9B Instruct at Q8_0, (a) Baseline method, (b) Think2 method



(a)

(b)

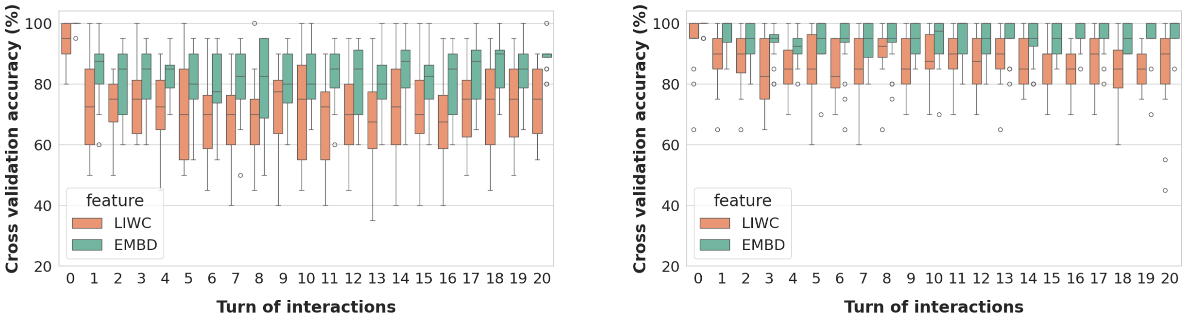
Figure 14: Cross validation accuracy of linguistic features from Gemma2 9B Instruct at Q4_0, (a) Baseline method, (b) Think2 method



(a)

(b)

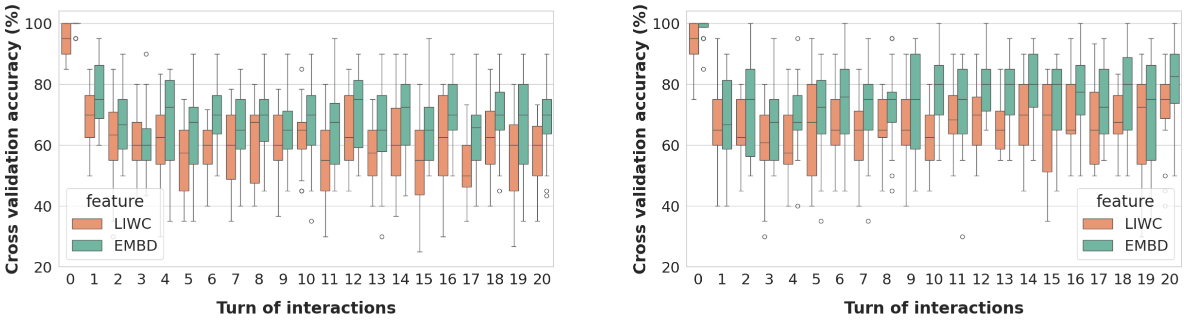
Figure 15: Cross validation accuracy of linguistic features from LLaMA3 8B Instruct at float16, (a) Baseline method, (b) Think2 method



(a)

(b)

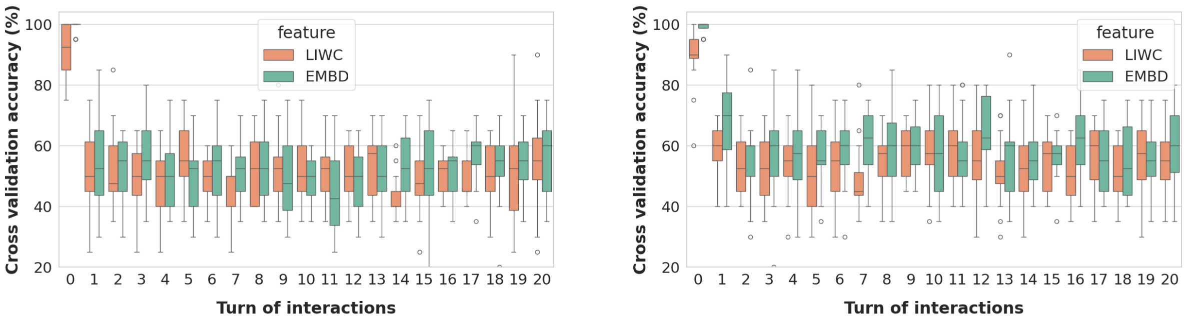
Figure 16: Cross validation accuracy of linguistic features from LLaMA3 8B Instruct at Q8_0, (a) Baseline method, (b) Think2 method



(a)

(b)

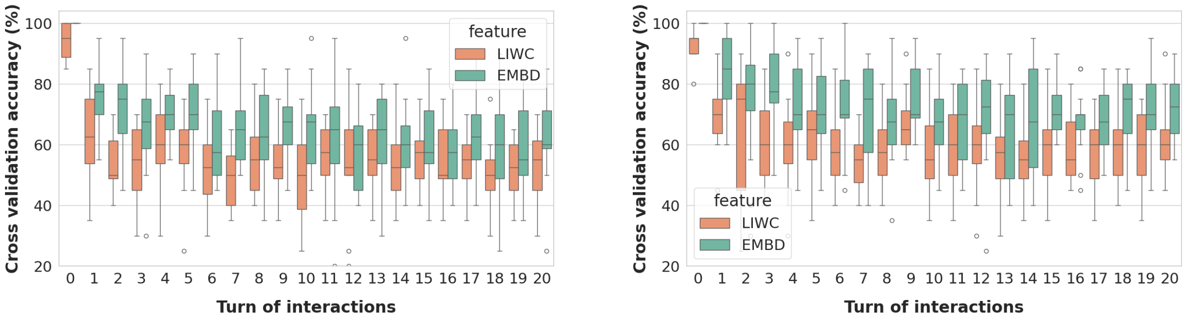
Figure 17: Cross validation accuracy of linguistic features from LLaMA3 8B Instruct at Q4_0, (a) Baseline method, (b) Think2 method



(a)

(b)

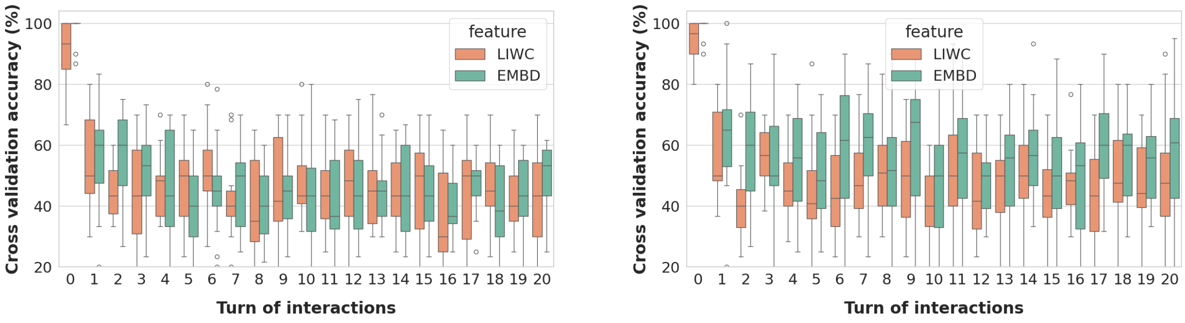
Figure 18: Cross validation accuracy of linguistic features from Mistral 7B Instruct v0.3 at float16, (a) Baseline method, (b) Think2 method



(a)

(b)

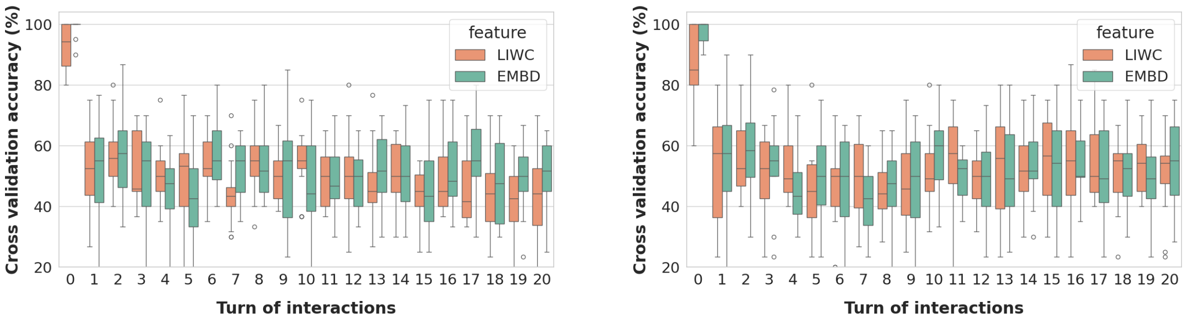
Figure 19: Cross validation accuracy of linguistic features from Mistral 7B Instruct v0.3 at Q8_0, (a) Baseline method, (b) Think2 method



(a)

(b)

Figure 20: Cross validation accuracy of linguistic features from Mistral 7B Instruct v0.3 at Q4_0, (a) Baseline method, (b) Think2 method



(a)

(b)

Figure 21: Cross validation accuracy of linguistic features from Gemma 7B Instruct v1.1 at float16, (a) Baseline method, (b) Think2 method

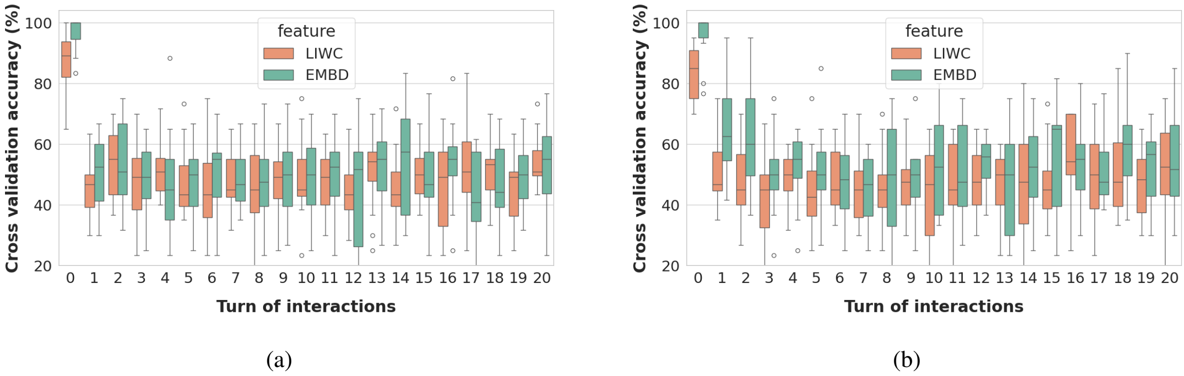


Figure 22: Cross validation accuracy of linguistic features from Gemma 7B Instruct v1.1 at Q8_0, (a) Baseline method, (b) Think2 method

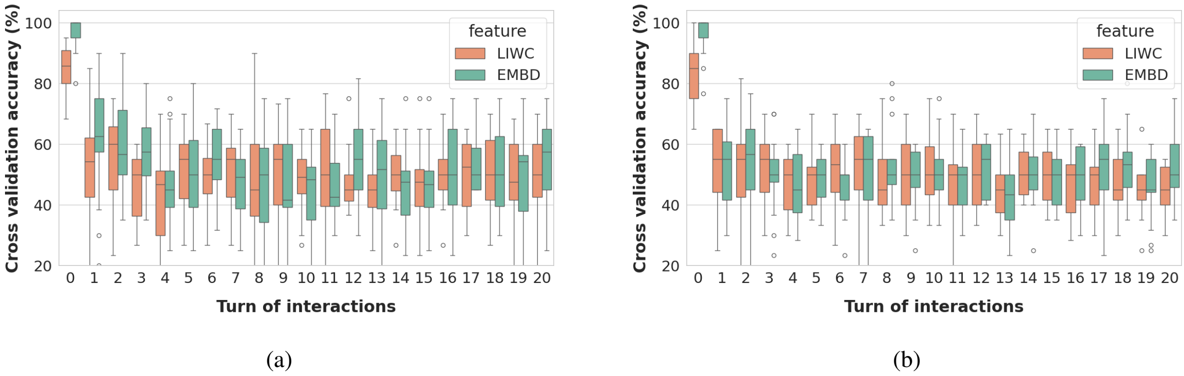
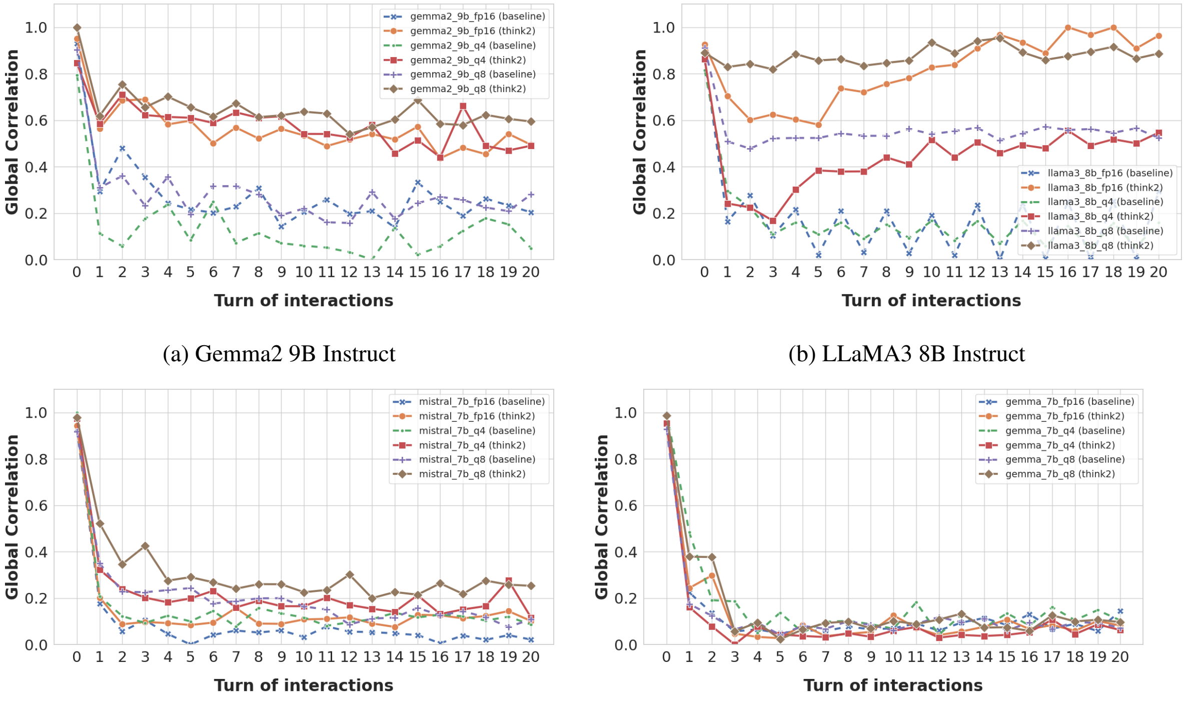


Figure 23: Cross validation accuracy of linguistic features from Gemma 7B Instruct v1.1 at Q4_0, (a) Baseline method, (b) Think2 method



(a) Gemma2 9B Instruct

(b) LLaMA3 8B Instruct

(c) Mistral 7B Instruct v0.3

(d) Gemma 7B Instruct v1.1

Figure 24: Global correlation plot at different quantization levels with Baseline and Think2 methods from (a) Gemma2 9B Instruct, (b) LLaMA3 8B Instruct, (c) Mistral 7B Instruct v0.3, (d) Gemma 7B Instruct v1.1