

Do Personality-Tuned LLMs Make Better Social Agents?

Tim Krabbe¹ and Xiaodan Shi¹

Abstract—LLMs are increasingly used in social simulations for socially interactive agents and robots, offering more flexibility than rule-based systems. However, even though they mimic human behaviour very well, there is a persistent *alienness* to them. This work investigates whether personality-aware fine-tuning can reduce this gap by improving the consistency and controllability of personality-conditioned dialogue generation compared with instruction prompting alone. We fine-tune two small open-weight LLMs, Qwen2.5-7B-Instruct and Ministral-8B-Instruct, using a corpus that combines personality-labelled social media posts and dialogues to create a personality-based dialogue engine for social simulation. The resulting models are evaluated across multiple social interaction scenarios using three independent LLM judges, which assess personality fidelity and provide evidence-based behavioral interpretations. We additionally quantify inter-rater agreement and lexical characteristics of the generated dialogue. Results indicate that fine-tuned models are not better at role-playing different personalities than their respective baseline models. However, low inter-rater agreement limits the confidence with which these results can be interpreted. Concerning the quality of generated texts, fine-tuned models are mostly comparable to the baselines, with fine-tuning improving the linguistic diversity of the Qwen models. While the results appear generally usable and the baseline models offer the best overall performance, future studies should place greater emphasis on the quality and domain alignment of training data for accurate personality role-playing.

I. INTRODUCTION

Large language models (LLMs) are increasingly used as behavioral models for socially interactive agents and robots. Their ability to generate context-dependent language and adapt to open-ended interaction offers a flexible alternative to conventional rule-based or utility-based agent models, particularly in applications involving human-robot interaction, social robotics, and multi-agent simulation [1]–[3]. However, linguistic fluency alone does not imply behaviorally consistent social interaction. LLM agents may produce responses that appear plausible at the surface level while lacking stable behavioral characteristics across contexts. Anthis et al. [4] describe this discrepancy as a form of *alienness*, arising from the mismatch between apparently human-like outputs and the non-human mechanisms that generate them.

For socially interactive robotic systems, this limitation directly affects the controllability and interpretability of agent behavior. A robot or simulated agent intended to represent a particular user type, interaction style, or social role should exhibit behavioral characteristics that remain distinguishable across different interaction scenarios. Personality provides one possible abstraction for imposing such consistency. Previous work has shown that LLM-based agents can gen-

erate diverse social behaviors [5], [6], but it remains unclear whether explicit personality-oriented fine-tuning leads to more consistent personality-conditioned behavior than prompting pretrained instruction-following models alone. This question is also relevant to the use of LLM agents as synthetic populations, or *silicon samples* [7], where useful simulation requires not only diverse outputs but controllable behavioral variation among agents.

In this work, we investigate whether personality-aware fine-tuning can mitigate the alienness of LLM-generated social behavior. We use the Myers-Briggs Type Indicator (MBTI) [8] as an operational personality-conditioning scheme and compare personality-fine-tuned models with their corresponding instruction-tuned baselines. Two open-weight LLMs, Qwen2.5-7B-Instruct and Ministral-8B-Instruct-2410, are fine-tuned using an augmented personality-labelled corpus. We focus on small open-weight LLMs because they are more practical for resource-constrained social-agent applications and easier to fine-tune and deploy. The corpus combines MBTI-labelled online forum posts from the Kaggle MBTI dataset [9] with conversational data from DailyDialog [10]. To obtain personality annotations for the dialogue data, we train RoBERTa-based classifiers for each of the four MBTI dimensions and use them to assign personality labels to DailyDialog utterances.

We evaluate the models in controlled social-interaction scenarios in which agents are instructed to generate dialogue conditioned on predefined personality profiles. To assess whether the generated behaviors reflect the intended personality conditions, we introduce a two-stage evaluation framework combining personality classification with qualitative behavioral analysis. Inspired by deductive qualitative content analysis [11], three independent LLM judges evaluate each generated dialogue, infer its personality characteristics, and provide evidence-based justifications. Following the LLMs-as-a-Judge paradigm [12], the inferred personality labels are compared with the intended conditions, while agreement among judges is measured to assess evaluation reliability.

Through this setup, we address the following research question: *Does personality-aware fine-tuning produce more consistent and controllable personality-conditioned behavior than instruction prompting alone?* Rather than treating personality labels as psychological ground truth, we use them as structured behavioral conditions for evaluating whether an LLM can maintain distinguishable interaction patterns across social scenarios.

The main contributions of this work are:

- We present a personality-aware fine-tuning framework (Fig. 1) for dialogue-capable social agents, combining

¹Department of Computer and Systems Sciences, Stockholm University, 164 55, Kista, Sweden. xiaodan.shi@dsv.su.se

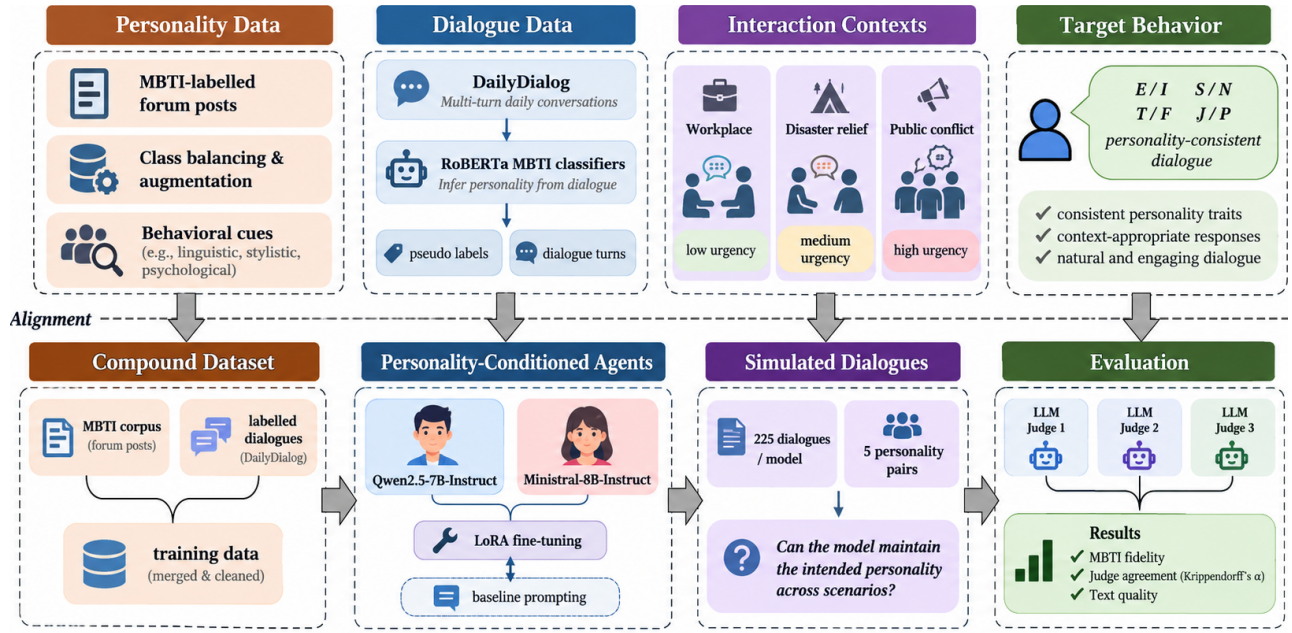


Fig. 1. Overview of personality-aware fine-tuning social simulation framework with LLMs. The top row shows the source components (personality data, dialogue data, interaction contexts, and target behavior), and the bottom row presents the aligned simulation pipeline (compound dataset, personality-conditioned agents, simulated dialogues, and evaluation).

personality-labelled natural language data with conversational data for training LLM-based behavioral models.

- We propose a multi-judge evaluation framework for measuring the consistency of personality-conditioned agent behavior across controlled social-interaction scenarios, incorporating both quantitative label agreement and qualitative behavioral evidence.
- We conduct a systematic comparison of pretrained and personality-fine-tuned variants of two open-weight LLMs, examining when fine-tuning improves the controllability and consistency of personality-conditioned social behavior.

II. RELATED WORK

A. LLMs for Social Simulation

How to leverage LLMs in the social sciences is a question that is naturally debated throughout many areas of the field. Ziems et al. [13] test various models for field-specific tasks and provide a range of potential use-cases. For social simulation, LLMs have the potential to replace game theoretical approaches by powering agents in a simulated environments [14]. Rule-based utility functions, as they are commonly used in simulation approaches, are quite limited in how realistic they can represent human behaviour. They are dependent on consistent preference orderings of the agent, which is a significant simplification of human decision-making.

So far, several studies have explored the possibility of using LLMs in various ways for social simulation. Gurcan [15] describes advantages and limitations of using AI for those simulations, for example accessibility, data collection and preparation, datafication or support in gathering insights from the simulation model. Most importantly though, a major

advantage of AI is described as the generation of agent-based models and scenarios as well as the explainability of the agents, i.e. the generation of explanations of actions, decisions and mechanics of the simulations in natural language. For data collection for example, [16] use language models to synthesise high-quality population-aligned persona sets for social simulation based on social media data, significantly reducing population-level bias. [17] use a language model to make their simulation framework more accessible to support framework and output explanations, generating visualisations and usage guidance.

B. Multi-Agent Simulation

A common technique for social simulation is agent-based modelling (ABM), which is used to simulate the actions and interactions of autonomous agents in a given system with the goal of mimicking social processes to enable exploration and understanding of social systems [15]. ABM can be broken down into three types of simulation: individual simulations, scenario simulation and society simulation [2]. For instance, LLMs show strong capabilities in conducting role-play, using either zero-shot prompting to simulate human behaviour and interactions [18], [19].

Full scenario or society simulation implementations are usually done in sandbox environments made specifically for LLM powered multi-agent simulation [20]. Some studies have even explored personality-driven behaviors in such simulations [21], [22], but the effect of personality-oriented fine-tuning on the consistency of such behaviour remains less explored.

Regarding dialogue simulation, [5] demonstrate the capabilities of GPT-3.5 to drive society simulations. Agents move

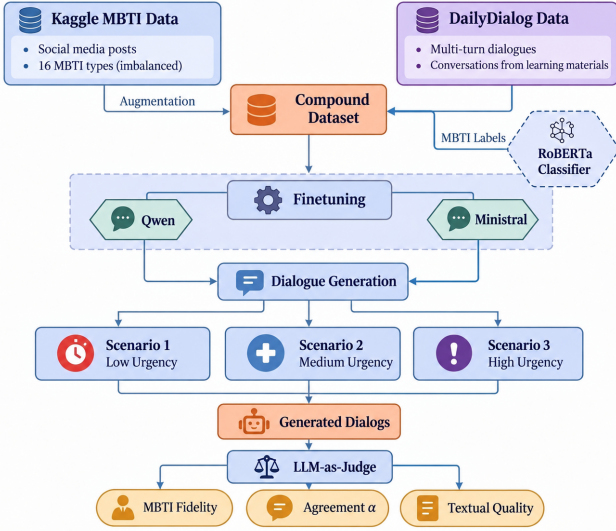


Fig. 2. Experimental Setup

through the *Smallville* society and interact with each other through user input and LLM-based generation of dialogues and autonomous actions derived from those. Another application of LLMs for dialogue generation for social simulation is shown in [6]. The authors use a Gemini model which is fed situational contexts, the social state and practice as well as the conversation history to generate fitting dialogue.

C. Personality

While there is no absolute, quantifiable way of measuring personality, there are several frameworks that are frequently applied to approximate it. The most common ones are the Myers-Briggs Personality Type (MBTI) model [8] and the Big Five personality type model [23]. The MBTI classification classifies personality along four dimensions: Extraversion-Introversion (I/E), Sensing-Intuition (N/S), Thinking-Feeling (F/T) and Judging-Perceiving (J/P). This totals to 16 different personality types. While the MBTI test is criticised in personality research regarding its reliability and validity [24] and the Big Five model is often preferred for clinical application and psychological research [25], it is still often used for text classification in computer science contexts [26]–[29]. Since the MBTI framework is made up of 4 binary dimensions compared to 5 continuous ones in the Big Five model, it is easier to use and to evaluate, which is why for this study the MBTI model was preferred.

III. METHOD

Fig. 2 gives an overview of the method of the study. Two different LLMs are fine-tuned on a compound dataset made up of MBTI-labelled social media posts and multi-turn dialogues, using two different LoRA-configurations. The four fine-tuned models and their untuned baselines are then repeatedly prompted to act as conversational agents to generate dialogues based on certain personality types taking

place in a range of exemplary scenarios. This results in six sets of 225 dialogues, one per model variant, which are then evaluated with a combination of qualitative and quantitative methods using an LLMs-as-a-Judge approach. MBTI fidelity of the generated texts, examples of the qualitative analysis, judge agreement and textual quality are then presented as results.

A. Personality-Tuned LLMs

Two models were chosen for the fine-tuning experiments: Qwen2.5-7B-Instruct [30] and Minstral-8B-Instruct-2410 [31]. The Qwen model is a lightweight model with 7.6 billion parameters, 28 layers and 28 attention heads, while the Minstral model has 8 billion parameters, 36 layers and 32 attention heads. Both come with a context window of 128k tokens. Their small size and relatively large context window make them interesting candidates for the underlying task, where compute might be limited but large amounts of text in form of conversational histories and memories need to be processed. Both models are distributed under research friendly licenses, Apache 2.0 and mrl respectively, which allow free use, modification, and redistribution for research purposes, ultimately facilitating improved reproducibility.

The models were fine-tuned using Low-Rank Adaptation (LoRA). LoRA is a method for fine-tuning, in which the weight update ΔW is decomposed into two trainable low-rank matrices B and A , encoding task-specific adaptations into a compact representation $\Theta = \{B, A\}$, while keeping W_0 frozen. Through this encoding, LoRA can reduce the number of trainable parameters from $d \times k$ to $r(d + k)$, where $r \ll \min(d, k)$. This enables fine-tuning with much less resources and on smaller hardware setups. Additionally, LoRA helps to prevent catastrophic forgetting, as the pre-trained weights remain frozen and only the low-rank adapter matrices are updated [32]. This is especially important for a small model when its pre-trained instruction following capacities need to be preserved.

LoRA is defined as

$$h = W_0 x + \frac{\alpha}{r} B A x \quad (1)$$

where h is the output, W_0 is the frozen pre-trained weight matrix, x is the input, α is a scaling constant, r is the rank, and B and A are the trainable low-rank decomposition matrices. As loss function of the model the default setting is used, which is the cross-entropy loss [30], [33]. It is defined as

$$\mathcal{L} = -\frac{1}{T} \sum_{t=1}^T \log P_{\theta}(x_t | x_{<t}) \quad (2)$$

where x_t is the token at step t , $x_{<t}$ represents the previous tokens or the context, and θ is the parameter vector of the model.

B. Personality-Awareness of LLMs

The study uses a two-step evaluation pipeline made up of qualitative textual content analysis and subsequent classification conducted with an LLMs-as-a-Judge approach. For this,



Fig. 3. Illustration of the three scenarios representing low urgency, medium urgency and high urgency.

a qualitative textual analysis using deductive coding based on the framework by Mayring [11] was conducted by three LLMs using an LLMs-as-a-Judge approach. Three scenarios were created, representing three different levels of urgency of the depicted social situations. The fine-tuned models and their respective baseline models were then prompted to generate a set of dialogues. For each dialogue, each turn equals one model call, in which the model is provided the personality type of the current speaker, the conversation history, the scenario and a brief set of generation rules. Five pairs of speaker personality types were applied. For each combination of personality pair and scenario a number of 15 dialogues were generated, totalling in 225 dialogues per model. An overview of the setup is displayed in Fig. 2.

1) *Simulation Scenarios*: Three scenarios (Fig. 3) were created to test the models capabilities over a range of settings. The scenarios differ in both the setting they are taking place in and the urgency of the conversing agents. The first scenario is a low urgency situation at the workplace: two agents are discussing which of the two should lead the next team meeting. The second scenario displays a medium urgency situation in the context of disaster relief. The two agents were given the task to organise the distribution of medical goods in a disaster-struck area. They were told they were among the first-responders to the situation and therefore had to operate under incomplete information. The scenario is defined as medium urgent relative to the situation: it is a planning situation that calls for a professional solution without immediate danger for either of the participants. The third scenario displays a residential setting: the two agents role-play neighbours who meet in the supermarket. A national health order was introduced that requires individuals in public indoor spaces to wear a mask, as it was the case in many societies during the Covid-19 pandemic. One agent complies with the order, while the other one does not. The dialogue should contain the confrontation of both agents, which is why the scenario is defined as highly urgent.

2) *LLMs-as-a-Judge Evaluation*: The evaluation of the generated dialogues was done by applying an LLMs-as-a-Judge approach, in which the judging models were few-shot prompted to provide a textual content analysis loosely based on the method developed by [11]: The judging models are asked to provide the marked text segment, the MBTI

Example: Generated Dialogue by Ministral-8B

Scenario: Project meeting

Person A: We have made many great strides during your tenure as head of our department. But we still have a long road ahead, especially if we continue doing things the same old way.

Person B: I know exactly where you're coming from. But we need someone with hands-on experience across all departments. There isn't much time left until the deadline. We need to make tough decisions here!

Person A: The people who sit around the table every day don't get things done either. If anything, I've found they slow us down most of the time. They talk themselves into oblivion instead of just doing their jobs.

Person B: These members are not exactly enthusiastic about change. So it's up to me to take charge of this meeting. I'm sure you see that, right? People think I act like a jerk because I speak plainly.

Mistral-Small-3.1 Judge Codings

A – Thinking (T): “They ... slow us down most of the time.”
Critical and objective assessment of others' performance.

A – Judging (J): “They talk themselves into oblivion instead of just doing their jobs.”
Values action and closure over discussion.

B – Sensing (S): “We need someone who has hands-on experience across all departments.”
Emphasizes practical, experiential knowledge.

B – Thinking (T): “People think I act like a jerk because I speak plainly.”
Values direct and objective communication.

Fig. 4. Example of a generated dialogue and selected personality codings produced by an LLM judge.

dimension they argue is signified by the segment and a short reasoning statement to justify this decision. They are then asked to propose a full MBTI label derived from these deductive codings. The prompt itself makes use of persona prompting, telling the model to role-play as a psychometrics expert and qualitative researcher. The different dimensions of the MBTI framework are briefly made clear with a range of fitting keywords. The input format of the generated dialogues is explained with an example as well as the expected output format, including an exemplary coding of two dialogue segments (Fig.4).

Accuracy and F1 scores are calculated to measure how

well the judges can detect the displayed MBTI types in the generated dialogues. The ability to do so is used as a proxy measure for how well the models can display the personality types in their outputs, i.e. if the judging models are assessing the personality types displayed in the generated dialogue well, this is interpreted as the generating model displaying this personality type convincingly. However, since this is a proxy measurement, conclusions will have to be drawn with caution. To carry out the qualitative analysis, three medium-sized models were chosen: Qwen3-32B, Qwen3-30B-A3B-Instruct-2507 [34] and Mistral-Small-3.1-24B-Instruct-2503 [35]. This way, diversity of the judging models is ensured: Mistral and Qwen represent different model families, while Qwen3-30B-A3B-Instruct-2507 is based on a mixture of experts architecture as opposed to the other two. All three are distributed under the Apache 2.0 license.

3) *Inter-Rater Reliability and Textual Quality*: Inter-rater reliability was assessed using Krippendorff’s α , which supports multiple raters and missing observations [36]. The MBTI dimensions were treated as nominal variables. McNemar’s test was used to compare paired classification outcomes between baseline and fine-tuned models. To analyse and compare the textual quality of the dialogue outputs, the English-rate, Distinct-1 and Distinct-2 were calculated.

IV. RESULTS

A. Dataset

To conduct the experiments, we selected the (MBTI) Myers-Briggs Personality Type Dataset [9] from Kaggle, which is available under the CC0 public domain license. The dataset contains originally 8675 observations, labelled with MBTI types, e.g. INTJ or ESFP. Each observation contains 50 online forum posts by the same person, which are split up into respective posts. Although the data represents conversational texts, it does not include dialogues. Because of this, for fine-tuning purposes the DailyDialog dataset [10] was added to the data. This dialogue data was labelled using a RoBERTa classifier trained on a class-balanced Kaggle dataset with equal representation of all 16 MBTI types. After the pre-processing steps, the dataset contained approximately 582,000 observations, the dialogue data making up a share of roughly 12.8%.

Due to the large class imbalances of the Kaggle dataset, augmentation was necessary. To achieve a more balanced class distribution, oversampling via backtranslation was applied for the most imbalanced class, the N/S dimension of the MBTI framework. This was done using the MarianMT model, translating text from English to German and back. In this way, the imbalance of the N/S dimension went from approximately 1:7 to 1:3.7. Augmentation quality was ensured using a number of different metrics: BLEU, Self-BLEU, Type Token Ratio, Cosine Similarity and Perplexity.

B. Experiment settings

Fine-Tuning Outcomes In all four instances, model training was ended by the early stopping callback triggering after reaching a plateau in the evaluation loss curve. For both

Qwen models this was the case after 1.004 epochs, which approximates to 7400 steps. Both models show a very similar

TABLE I
FINE-TUNING HYPERPARAMETERS AND LoRA CONFIGURATIONS

Hyperparameter	Value
<i>LoRA Configurations</i>	
LoRA 1 (Rank r / Alpha α)	16 / 32
LoRA 2 (Rank r / Alpha α)	32 / 64
Dropout	0.05
Target Modules	q_proj, k_proj, v_proj, o_proj
<i>Training Parameters</i>	
Learning Rate	2×10^{-5}
Warmup Ratio	0.03
Optimizer	AdamW
Batch Size (per device)	8
Gradient Accumulation Steps	2
Number of Epochs	6
Early Stopping Patience	5
Max Sequence Length	512
Mixed Precision	bfloat16

evaluation loss of 1.293 (LoRA: $r = 16, \alpha = 32$) and 1.289, and a evaluation mean token accuracy of about 0.731. For the first Ministal model ($r = 16, \alpha = 32$), training was ended after approximately 1.574 epochs and for the second one ($r = 32, \alpha = 64$) after 2.117 epochs. The first model reaches an evaluation loss of 1.99 with mean token accuracy of 0.594 and the second one 1.978 with a mean token accuracy of 0.614. The hyperparameters for the fine-tuning can be seen in I. The parameters here are chosen following the findings by [32], which indicate that targeting more weight matrices generally improves performance and thus the LoRA adapters were applied to the query, value, key and output projections of all transformer layers. For the Qwen models, this resulted in roughly 10.09 (0.13%) respectively 20.19 million (0.27%) trainable parameters, and for the Ministal models 15.34 (0.19%) and 30.67 million (0.38%) trainable parameters.

Evaluation Metrics To evaluate personality fidelity, accuracy and macro F1 were calculated for each of the four MBTI dimensions, measuring the agreement between the intended personality conditions and the labels inferred by the judging models. Inter-rater reliability was assessed using Krippendorff’s α , quantifying the agreement between judges. To assess linguistic characteristics, the proportion of English-language output was measured, while lexical diversity was evaluated using Distinct-1 and Distinct-2, representing the proportion of unique unigrams and bigrams, respectively.

C. Qualitative Evaluation and Inter-Rater Agreement

The LLMs-as-a-Judge approach has produced a large amount of qualitative data and deductive codings, which would be out of scope for this article to analyse in great detail. However, a number of exemplary findings should be demonstrated. Fig. 5 shows how the judging models at times focus on different things about the same text segment, coming to different conclusions: While both Qwen models reasoned that the speaker was reflecting on a hypothetical

scenario, which they interpreted as a marker for *Intuition (N)*, the Mistral model instead focused on the aspect that sparked this reflection, which lies in the present and the need to wear a mask, labelling the segment with *Sensing (S)*.

Utterance (Person A):

“If someone had told me six months ago that we’d have to wear masks at all times, I wouldn’t believe them.”

Judge 1 (Qwen3-32B) → Intuition (N)

Reasoning: Focuses on future possibilities and abstract scenarios.

Judge 2 (Qwen3-30B) → Intuition (N)

Reasoning: Reflects on a hypothetical scenario contradicting past expectations.

Judge 3 (Mistral-Small-24B) → Sensing (S)

Reasoning: Focuses on the present and practical aspects of mask-wearing.

Fig. 5. Example of inter-rater disagreement caused by differing dimensional interpretations of an identical text segment.

Strong agreement could also be explained by very similar reasoning: For example, the sentence “Hey! Put that thing back up!” is described by every judge as *direct* and *action-oriented*, resulting in a unison *Extraversion (E)* label. At other times models might get conflicting information about the speaking person, for example, when at first strong criticism is uttered (“You really need to stop making excuses for yourself”, *Thinking (T)*), but then an affectionate goodbye follows (“Take care, by! Love you lots!”, *Feeling (F)*). The latter also giving a hint on a strong limitation of the used dataset: social media posts being written in a certain type of language, which might make the generated utterance fall out of place. A goodbye like that would be much more common on a social media platform than in a work context.

The inter-rater reliability was calculated by applying Krippendorff’s Alpha. Since the Judging Models decided for themselves which text segments to code, they haven’t always coded the same ones. For when they have, the agreement looks as shown in table II. Agreement is highest on the J/P dimension, reaching the recommended threshold of 0.800. Agreement on the N/S dimension reaches the 0.667 threshold that allows to draw what Krippendorff calls “tentative conclusions” [36], and the other two dimensions miss it slightly. However, the small sample size for any other dimension than I/E means that these results need to be viewed with caution.

Agreement on the final judgement of the speakers is shown in table III. A few things stand out: first of all, agreement on the text of the baseline models was higher in the F/T and the J/P dimension, and also in the N/S dimension for the Mistral baseline. However, for the I/E dimension and N/S for the Qwen model family, agreement was highest for the model with the smaller LoRA configuration. Second, the α value for the N/S dimension is immensely lower throughout all models than in any other dimension. Lastly, the threshold of $\alpha = 0.8$ was not reached for any model or dimension, making it fairly difficult to draw robust conclusions from

these results.

TABLE II
INTER-RATER RELIABILITY (KRIPPENDORFF’S α) EVALUATED ON IDENTICAL TEXT SEGMENTS.

Dimension	α	Identical Segments (n)
I/E	0.644	57
N/S	0.693	10
F/T	0.600	3
J/P	0.847	9

D. MBTI Fidelity

The evaluation is conducted on a test set of $N = 225$ dialogues per generating model. Since each multi-turn dialogue features two interacting personas (Person A and Person B), this corresponds to a maximum of $N = 450$ individual persona evaluations per model. Due to missing judge outputs for some dialogues, the actual number of evaluated persona instances ranges from 392 to 450 across models.

Table III shows that across both model families and all four MBTI dimensions, the judges generally achieved the highest classification performance on the outputs generated by baseline models. Applying the interpretation of this measurement as a proxy for the capability of the model to role-play certain personality types, this result shows that fine-tuning does not yield a better performance in doing so compared to the baseline models.

Depending on the dimension, F1 scores for the baseline models fall between 0.595 (I/E dimension) and 0.820 (F/T) for Mistral and between 0.530 (I/E) and 0.791 (F/T) for Qwen. For both model families, judges were most accurate at classifying the F/T dimension with, followed by the J/P, the N/S and lastly the I/E dimension. This pattern also holds for the fine-tuned models. The F1 scores for the fine-tuned models were generally lower, ranging from 0.446 (LoRA: $r = 16, \alpha = 32$) to 0.738 ($r = 16, \alpha = 32$) for Mistral and from 0.425 ($r = 32, \alpha = 64$) to 0.777 ($r = 16, \alpha = 32$) for Qwen.

If higher rank r and increased scaling factor α make it easier for the judging models to detect the displayed personality type is not conclusively inferable from the results. Performance decreased slightly on the I/E and F/T dimensions for both model families, while on N/S it showed small improvements. Results for J/P differed between the two model families, with an improvement for Mistral but a slight decrease for Qwen.

Calculating McNemar’s Test to test for significant differences between the model configurations shows that the outputs of baseline models for both model families are significantly more accurately classified in accordance with the supposedly role-played personality, than those generated by the fine-tuned models. Only the F/T dimension in the texts generated by the Qwen model with the smaller LoRA configuration was not significantly more accurately classified by the judges than its baseline.

TABLE III

COMPREHENSIVE EVALUATION RESULTS: CONSENSUS FIDELITY (ACCURACY, MACRO F1), INTER-RATER RELIABILITY (KRIPPENDORFF’S α), AND LINGUISTIC GENERATION QUALITY (ENGLISH PROPORTION, DISTINCT-1/2).

Model Variant	Fidelity (Acc, F1) & Reliability (α)												Linguistics		
	I/E			N/S			F/T			J/P			Eng (%)	D-1	D-2
	Acc	F1	α	Acc	F1	α	Acc	F1	α	Acc	F1	α			
Ministral															
Base	0.596	0.595	0.435	0.620	0.607	0.164	0.820	0.820	0.621	0.760	0.758	0.654	99.97	0.0889	0.5450
LoRA ($r = 16, \alpha = 32$)	0.462	0.462	0.452	0.500	0.486	0.086	0.738	0.738	0.453	0.633	0.633	0.382	99.56	0.0882	0.5160
LoRA ($r = 32, \alpha = 64$)	0.449	0.446	0.425	0.524	0.517	0.080	0.729	0.729	0.487	0.687	0.687	0.483	99.43	0.0968	0.5383
Qwen															
Base	0.533	0.530	0.428	0.604	0.596	0.024	0.791	0.791	0.691	0.729	0.721	0.583	88.35	0.0812	0.4268
LoRA ($r = 16, \alpha = 32$)	0.458	0.455	0.555	0.529	0.525	0.045	0.778	0.777	0.540	0.673	0.672	0.457	92.03	0.0961	0.5524
LoRA ($r = 32, \alpha = 64$)	0.429	0.425	0.406	0.549	0.543	0.027	0.758	0.758	0.523	0.669	0.668	0.393	91.71	0.0965	0.5450

TABLE IV

MEAN MACRO F1 ACROSS THE FOUR MBTI DIMENSIONS FOR EACH SCENARIO (LOW, MEDIUM, AND HIGH URGENCY).

Model Variant	Low U.	Medium U.	High U.
Ministral			
Base	0.711	0.690	0.681
LoRA ($r = 16, \alpha = 32$)	0.558	0.574	0.605
LoRA ($r = 32, \alpha = 64$)	0.555	0.602	0.626
Qwen			
Base	0.678	0.627	0.663
LoRA ($r = 16, \alpha = 32$)	0.573	0.618	0.628
LoRA ($r = 32, \alpha = 64$)	0.588	0.586	0.617

To examine whether personality fidelity varies with the social context, we additionally compare performance across the three scenarios, shown in table. IV. Mean Macro F1 across the four MBTI dimensions shows no consistent monotonic relationship with scenario complexity. While some dimensions vary substantially between scenarios, the direction of these changes differs across models and dimensions. For example, F/T remains comparatively robust across all scenarios, whereas E/I shows consistently lower and more variable performance. Overall, these results suggest that scenario context affects the expression and recognition of individual personality dimensions, but does not systematically determine overall personality fidelity.

E. Linguistic Characteristics

To investigate linguistic characteristics of the generated dialogues, the rate of English-language text was calculated using the *langdetect* package for Python [37] and is shown in table III. The Ministral models consistently showed an English-rate of $> 99\%$. However, of the texts generated by the Qwen baseline model only 88.35% were in English or were detected as such; the majority the non-English texts being in Chinese. Fine-tuning raises the English-rate of the Qwen models to 92% for the smaller LoRA configuration respectively 91.7% for the larger one.

Distinct-1, measuring distinct unigrams, and Distinct-2, measuring distinct bigrams, were calculated to compare lexical diversity of the generated text, shown in table III. For Ministral, while the large LoRA configuration boosts the

Distinct-1 value by about 0.008 compared to the baseline and the small configuration, the baseline produces more diverse bigrams compared to the fine-tuned models. For the Qwen models, LoRA increases both values by about 0.015 for Distinct-1 and by about 0.125 for Distinct-2, with the smaller configuration producing the most diverse texts.

V. DISCUSSION

The results show that the judging models are better at classifying the texts that were generated by the baseline models than those of the fine-tuned models. Following the logic of this being a proxy measure for the capabilities of the generating models at role-playing different personalities, this implies that the proposed fine-tuning approach does not increase these capabilities compared to the baseline models. This result holds when compared across scenarios. Between the two applied LoRA configurations, the effects were ambiguous and neither could lift the performance above baseline. Judge agreement mostly lingers around the lower α threshold. Only in the N/S dimension agreement in the fidelity analysis showed a strong drop, compared to an unsuspecting α for identical text segments. This points to a general difficulty of the judging models to conclusively classify a speaker with either *Intuition* or *Sensing* characteristic, and possibly to mixed signals in the dialogues.

Using a dataset made up of social media posts for dialogue generation is a major domain shift, even though the posts are of conversational nature. Employing a labelled version of the DailyDialog dataset might not be sufficient to counter this domain shift. The fine-tuning approach of only varying the LoRA parameters *rank* and α but not also investigating different configurations of target modules, weakens the conclusion that fine-tuning did not improve personality role-playing performance and would have to be the subject of further studies.

Lastly, LLMs-as-a-Judge evaluation is subject to model-specific biases and stochasticity [38]. The low inter-rater agreement observed here highlights that personality-related qualitative judgments are particularly sensitive to evaluator interpretation. Increasing the number and diversity of judges and validating selected codings with human experts are promising directions for future work.

REFERENCES

- [1] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu, “Large language models for robotics: A survey,” 2023, unpublished.
- [2] X. Mou, X. Ding, Q. He, L. Wang, J. Liang, X. Zhang, L. Sun, J. Lin, J. Zhou, X. Huang, and Z. Wei, “From individual to society: A survey on social simulation driven by large language model-based agents,” 2024, unpublished.
- [3] P. Taillandier, J. D. Zucker, A. Grignard, B. Gaudou, N. Q. Huynh, H. Kong, and A. Drogoul, “From the fluency fallacy to the micro-to-macro validity gap: Opportunities and pitfalls of LLMs in social simulation,” 2026, unpublished.
- [4] J. R. Anthis, R. Liu, S. M. Richardson, A. C. Kozlowski, B. Koch, E. Brynjolfsson, J. Evans, and M. S. Bernstein, “Position: LLM social simulations are a promising research method,” in *Proceedings of the 42nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 267. PMLR, 2025, pp. 81 005–81 034.
- [5] J. S. Park, J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative agents: Interactive simulacra of human behavior,” in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. San Francisco, CA, USA: ACM, 2023, pp. 1–22.
- [6] M. Treanor, B. Samuel, and M. J. Nelson, “Prototyping slice of life: Social physics with symbolically grounded LLM-based generative dialogue,” in *Proceedings of the 19th International Conference on the Foundations of Digital Games*. Worcester, MA, USA: ACM, 2024, pp. 1–4.
- [7] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, “Out of one, many: Using language models to simulate human samples,” *Political Analysis*, vol. 31, no. 3, pp. 337–351, 2023.
- [8] I. B. Myers, *The Myers-Briggs Type Indicator: Manual*. Palo Alto, CA, USA: Consulting Psychologists Press, 1962.
- [9] “Myers-Briggs Personality Type Dataset,” Kaggle, dataset. [Online]. Available: <https://www.kaggle.com/datasets/datasnaek/mbti-type>
- [10] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset,” in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, G. Kondrak and T. Watanabe, Eds. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 986–995. [Online]. Available: <https://aclanthology.org/I17-1099/>
- [11] P. Mayring and T. Fenzl, “Qualitative inhaltsanalyse,” in *Handbuch Methoden der empirischen Sozialforschung*, N. Baur and J. Blasius, Eds. Wiesbaden: Springer Fachmedien, 2019, pp. 633–648.
- [12] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 46 595–46 623, 2023.
- [13] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, “Can large language models transform computational social science?” *Computational Linguistics*, vol. 50, no. 1, pp. 237–291, 2024.
- [14] C. A. Bail, “Can generative AI improve social science?” 2023, unpublished.
- [15] O. Gurcan, “LLM-augmented agent-based modelling for social simulations: Challenges and opportunities,” 2024, unpublished.
- [16] Z. Hu, J. Lian, Z. Xiao, M. Xiong, Y. Lei, T. Wang, K. Ding, Z. Xiao, N. J. Yuan, and X. Xie, “Population-aligned persona generation for LLM-based social simulation,” 2025, unpublished.
- [17] P. J. Giabbanelli, “GPT-based models meet simulation: How to efficiently use large-scale pre-trained language models across simulation tasks,” 2023, unpublished.
- [18] A. Kong, S. Zhao, H. Chen, Q. Li, Y. Qin, R. Sun, X. Zhou, E. Wang, and X. Dong, “Better zero-shot reasoning with role-play prompting,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2024, pp. 4099–4113.
- [19] M. Shanahan, K. McDonnell, and L. Reynolds, “Role play with large language models,” *Nature*, vol. 623, no. 7987, pp. 493–498, 2023.
- [20] J. Lin, H. Zhao, A. Zhang, Y. Wu, H. Ping, and Q. Chen, “AgentSims: An open-source sandbox for large language model evaluation,” 2023, unpublished.
- [21] A. N. N. Rende, T. Yilmaz, and Ö. Ulusoy, “Negotiating comfort: Simulating personality-driven LLM agents in shared residential social networks,” 2025, unpublished.
- [22] P. Hardy, L. S. Marcolino, and J. F. Fontanari, “The paradox of productivity during quarantine: an agent-based simulation,” *The European Physical Journal B*, vol. 94, no. 1, p. 40, Jan. 2021. [Online]. Available: <https://doi.org/10.1140/epjb/s10051-020-00016-4>
- [23] R. R. McCrae and O. P. John, “An introduction to the five-factor model and its applications,” *Journal of Personality*, vol. 60, no. 2, pp. 175–215, 1992.
- [24] D. J. Pittenger, “The utility of the Myers-Briggs Type Indicator,” *Review of Educational Research*, vol. 63, no. 4, pp. 467–488, 1993.
- [25] A. Furnham, “The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality,” *Personality and Individual Differences*, vol. 21, no. 2, pp. 303–307, Aug. 1996. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/0191886996000335>
- [26] M. Gjurković and J. Šnajder, “Reddit: A gold mine for personality prediction,” in *Proceedings of the Second Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media*. New Orleans, LA, USA: Association for Computational Linguistics, 2018, pp. 87–97.
- [27] S. Bharadwaj, S. Sridhar, R. Choudhary, and R. Srinath, “Persona traits identification based on Myers-Briggs Type Indicator (MBTI)—a text classification approach,” in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018, pp. 1076–1082.
- [28] S. S. Keh and I.-T. Cheng, “Myers-Briggs personality classification and personality-specific language generation using pre-trained language models,” 2019, unpublished.
- [29] M. H. Amirhosseini and H. Kazemian, “Machine learning approach to personality type prediction based on the Myers-Briggs Type Indicator,” *Multimodal Technologies and Interaction*, vol. 4, no. 1, p. 9, 2020.
- [30] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, “Qwen technical report,” 2023, unpublished.
- [31] Mistral AI, “Minstral-8b-instruct-2410,” Hugging Face, 2024, [Online]. Accessed Sep. 14, 2026. [Online]. Available: <https://huggingface.co/mistralai/Minstral-8B-Instruct-2410>
- [32] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” *International Conference on Learning Representations*, 2022, unpublished.
- [33] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.
- [34] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, L. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu, “Qwen3 technical report,” 2025, unpublished.
- [35] Mistral AI, “Mistral-small-3.1-24b-instruct-2503,” Hugging Face, 2025, [Online]. Accessed Sep. 14, 2026. [Online]. Available: <https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503>
- [36] K. Krippendorff, “Computing Krippendorff’s alpha-reliability,” 2011, unpublished.
- [37] M. Danilak, “langdetect,” GitHub, 2021, [Online]. Accessed Sep. 14, 2026. [Online]. Available: <https://github.com/Mimino666/langdetect>
- [38] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Z. Lin, B. Zhang, L. Ni, W. Gao, Y. Wang, and J. Guo, “A survey on LLM-as-a-judge,” *The Innovation*, vol. 7, no. 6, p. 101253, 2026.