# Extended Abstract

**Motivation** Research has shown that aligning LLMs with specific personality traits can improve their performance on certain tasks. For example, tasks requiring creativity or problem-solving may benefit from traits like openness to experience, while tasks requiring attention to detail may benefit from conscientiousness

Given this context, this project has the goal of proposing a methodology to generate agents aligned with personality traits described by the five big personality factors using the TinyTroupe framework Salem et al. (2024).

**Method** The methodology has been divided in three part: Data preparation, model training, run simulation with tiny troupe and agent evaluation.

- Data preparation: Real responses from the IPIP-120 personality assessment were used to generate agent specifications. These agent specifications served as the foundation for the TinyTroupe simulation and they were generated by prompting large and small language models. To construct the SFT training set, we used all prompt + agent specification generated in the previous step. To construct the DPO training set, agent specifications generated by each model were evaluated and ranked (reject or choosen agent specification) by GPT-4o.
- Model Training: The model was trained sequentially through SFT and DPO.
- TinyTroupe Simulation: Using the agent specifications, the TinyTroupe pipeline developed in this project was employed to simulate agents answering the IPIP-120 test.
- Agent Evaluation: After running the TinyTroupe simulation, the agents' responses were compared to the original human records from which the agent specifications were derived. The metrics used were: agent average accuracy, agent average score alignment and traits average accuracy.

**Implementation** The full implementation can be found here: Project Git

The dataset can be found here: Data

**Results**

- **Average Score Alignment:** For each personality trait, the model's responses and the individual's responses to a set of multiple-choice questions were compared. This metric measures the average difference between the model's and the individual's answers, with a lower score indicating greater alignment. A perfect alignment corresponds to a score of zero, and scores can range from 0 to 4. The score alignment for each agent was calculated and then averaged across all agents to obtain this metric.
- **Average Accuracy:** This metric represents the mean accuracy of the agents' responses when compared to the corresponding individuals' records used to generate each agent's specification.
- **Personality Traits Accuracy:** This metric is the average accuracy for each personality trait, aggregated across the 20 test agents.

**Discussion** Qwen model consistently achieved the highest accuracy for personality traits. Since the primary goal of this study is to align agent specifications with personality traits, these results suggest that the trained Qwen model is the most effective among those evaluated.

While the results suggest it might be possible to improve the alignment between agents' descriptions and personality traits, the work of Zhu et al. (2025), which uses prompts instead of persona-based agents, their alignment scores per trait suggest potentially better performance.

**Conclusion** In summary, our evaluation across 20 diverse agents demonstrates that the trained Qwen model achieves the highest alignment between agent responses and human personality traits in comparison with the baselines of this project. However The available data do not allow for strong conclusions regarding the benefits of model alignment for persona agent-based systems.

# Align Small Language Models for Personality-Consistent Agent Simulation

**Caroline Santos Marques da Silva**
Department of Computer Science
Stanford University
cm199204@stanford.edu

## Abstract

Recent research suggests that aligning large language models (LLMs) with specific personality traits can enhance their performance on various tasks. This project proposes a methodology to generate agents aligned with the Big Five personality factors using the TinyTroupe framework. The approach involves four key steps: preparing data by generating agent specifications from real IPIP-120 personality assessment responses using both large and small language models; training models sequentially with Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO); simulating agent behavior via the TinyTroupe pipeline; and evaluating agents by comparing their simulated responses to the original human records. Results show that the trained model consistently achieves the highest alignment between agent and human responses. While these findings indicate potential for improved alignment, current data limitations prevent strong conclusions Overall, this study demonstrates promising initial results but highlights the need for further research and larger datasets to fully assess the effectiveness of personality-aligned agents.

## 1 Introduction

The integration of personality traits into large language models (LLMs) has emerged as a significant area of research, driven by the potential to enhance user experience, improve task performance, and enable more personalized interactions.

Research has shown that aligning LLMs with specific personality traits can improve their performance on certain tasks. For example, tasks requiring creativity or problem-solving may benefit from traits like openness to experience, while tasks requiring attention to detail may benefit from conscientiousness

Given this context, this project has the goal of proposing a methodology to generate agents aligned with personality traits described by the five big personality factors using the TinyTroupe framework Salem et al. (2024).

One advantage of using agents is that they can interact with each other to solve tasks, and combining multiple personality traits may lead to more interesting and effective problem-solving than relying on a single personality trait.

## 2 Related Work

According to Zhu et al. (2024) results, the incorporation of specific personality traits into prompts can effectively improve the execution of specialized professional tasks by LLM, reflecting a closer alignment with human behavior.

Chen et al. (2024) investigated methods for controlling the personality of large language models (LLMs). They proposed a hierarchy of techniques for effective personality control, finding that Prompt Induction following Supervised Fine-Tuning (PISF) achieved the best results. This approach leverages the advantages of both supervised fine-tuning and prompt induction. Their study compared PISF with reinforcement learning from human feedback (RLHF) using PPO for model alignment, but did not include a comparison with Direct Preference Optimization (DPO).

Zhu et al. (2025) raised concerns about traditional alignment methods, which typically emphasize broad societal values and general behaviors. They argue that this one-size-fits-all approach may not be adequate for capturing the nuanced and personalized aspects of individual behavior.

Zhu et al. (2025) proposes a new method: Personality-Activation Search (PAS), a method that fine-tunes model activations to closely match human preferences. PAS identifies key activations within the model that significantly impact Personality and personalized behaviors and adjusts these activations toward specific preferences. Their results have shown that PAS method outperform ICL methods and common RL algorithm (DPO and PPO) for model preference alignment.

However, previous studies did not investigate the use of agents enhanced with comprehensive persona descriptions, including factors such as profession, cognitive states, and memory. For instance, Zhu et al. (2025) relied on a simple prompt to represent the personality trait of an LLM. (An example can be found here: personality_prompt.json)

Therefore, this project proposes to fine-tune a small language model (SLM) using Direct Preference Optimization (DPO) to improve the quality and alignment of simulated personalities within TinyTroupe, an experimental Python library for creating human-like agents.

## 3 Method

In this section, the methodology for generating and evaluating agent specifications, which serve as blueprints for personality-driven simulation within the TinyTroupe framework, is described. Each agent specification is designed to represent an individual based on their responses to the IPIP-120 personality test, with the aim of improving alignment between personality traits and agent behavior. To achieve this, records from the IPIP-120 dataset are used as input for a language model, which generates structured agent profiles. These profiles are then used to simulate responses to the IPIP-120 test, enabling direct evaluation of personality alignment.

A pipeline has been developed in which personality profiles are systematically constructed, agent specifications are generated using large and small language models, and data are prepared for supervised fine-tuning (SFT) and Direct Preference Optimization (DPO)1. A small model is subsequently fine-tuned in order to improve the generated agents4. In the following subsections, each step of the process is described in detail—from personality profile construction and agent specification, through data preparation and model training, to the evaluation protocol used to assess personality alignment.

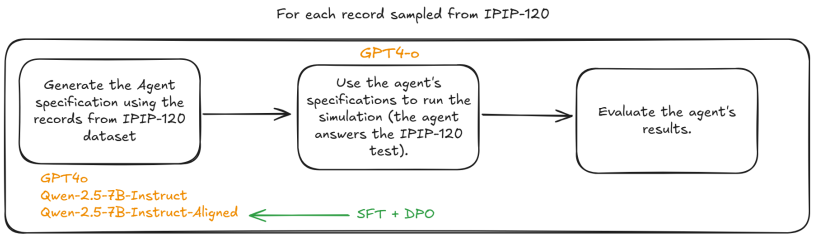


Figure 1: Agent Simulation Pipeline.

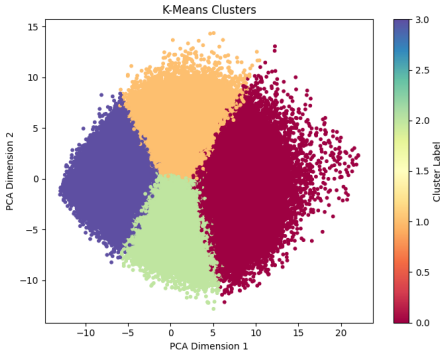


Figure 3: Personality Traits Clusters


Figure 2: Model Training Pipeline.

# 4 Experimental Setup

## 4.1 Personality Profile Contruction

The IPIP-NEO-120(Johnson, 2014) dataset, based on the Big Five personality model, was used as the foundation. Each record consisted of human responses to 120 items, mapped to 30 facets across five traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

To contextualize personality scores, individuals were grouped by sex and age, and percentiles (10th, 30th, 70th, and 90th) were computed within each demographic cohort. Cohorts were defined by combining sex with age groups (Teen, Young, Adult, Older Adult, Senior), such as 'Young Female' or 'Adult Male'. Each individual's facet and domain scores were then mapped to qualitative levels: Very Low, Low, Medium, High, or Very High—based on these percentile thresholds. These levels served as interpretable markers of trait intensity and were used as inputs for the prompt to generate the agent specification.

Sampling training and test set Given the large number of records in the IPIP-120 dataset, clusters were created using K-means, with questionnaire responses, sex, age, and country from each individual used as features. The resulting clusters are depicted in the figure below. To ensure consistent variability in the training and test sets, uniform sampling was performed from each cluster.

## 4.2 Agents Specification preparation

The `Qwen/Qwen2.5-72B-Instruct-Turbo`, `Qwen/Qwen2.5-7B-Instruct-Turbo` and `Qwen/QwQ-32B` models hosted by Together AI was used to generate these agent specifications. To encourage diversity in personality expression and exploration in the model inference space, the temperature parameter was varied in range of $+10\%$ and $-10\%$ during the process of data generation for each model.

Using the structured personality profiles, detailed prompts were constructed, including trait levels, facet levels, age, sex, and country.

The model outputs a structured JSON object that includes fields such as name, age, gender, nationality, occupation, long-term goals, beliefs, preferences, behaviors (daily routines), and a nested personality section that mirrors the Big Five trait structure. These specifications serve as the blueprint for agent

behavior in the TinyTroupe framework(Salem et al., 2024), enabling the simulation of personality-aligned human-like agents.

### 4.3 DPO and SFT Dataset preparation

It is important to say that The model `Qwen/Qwen2.5-7B-Instruct-Turbo` is not available on hugging face, thus, we use `Qwen/Qwen2.5-7B-Instruct` for model training.

Model fine-tuning is performed in two phases. First, in the first phase, a supervised fine-tuning (SFT) dataset was constructed from labeled prompts (prompt + agent spec). This phase helps the model learn to produce well-structured and coherent JSON outputs.

In the second phase, Direct Preference Optimization (DPO) is applied. For each personality prompt, two agent specifications were generated using different temperature settings. These are then evaluated using GPT-4o, which is prompted to judge which agent better matches the original personality description. The evaluation is based on trait-level fidelity, behavioral consistency, realism, and demographic alignment. The result is a preference-labeled dataset of approximately 3000 comparison pairs, which serves as the foundation for DPO fine-tuning. The models compared were as follows:

- `Qwen/Qwen2.5-72B-Instruct-Turbo` and `Qwen/QwQ-32B`
- `Qwen/Qwen2.5-72B-Instruct-Turbo` and `Qwen/Qwen2.5-7B-Instruct-Turbo`

At the end of the pipeline, the model `Trained-Qwen` was generated.

### 4.4 Agents Evaluation

Alignment was evaluated by simulating IPIP-120 responses from each generated agent and comparing them to the original responses used to generate that agent. Accuracy was computed at the item level, measuring the agreement between each agent's answer and the corresponding original human answer to the same IPIP-120 question as well as the accuracy with the pesonality trait level results. Additionally, the score alignment were computed.For each personality trait, we compare the model's and the person's answers to a set of multiple-choice questions, scoring how far apart their answers are on average. A lower score means the model's responses are more aligned with the individual's behavior, with a perfect alignment scoring zero. The range for score is from 0 to 4.

Due to resource limitations, this evaluation was conducted on a representative sample of 20 agents. In additional, running the TinyTroupe framework at scale, particularly for evaluating more than ten agents across multiple iterations is financially expensive. The inclusion of agents' episodic memory (previous test answers) in each iteration caused the prompt size to grow substantially which increases the costs.

### 4.5 Model Training

The training pipeline consisted of two stages: supervised fine-tuning followed by preference-based optimization. During the SFT phase, the model was trained on prompt–response pairs where the response was a well-formed agent specification. This allowed the model to learn to map psychological and demographic input into coherent personality-based agent descriptions.

After that, DPO was applied using the pairwise preference dataset. Each training sample included a personality prompt and two candidate agent specifications, with a label indicating the preferred version. The DPO objective adjusted the model to increase the likelihood of generating outputs more aligned with the labeled preferences. This two-step process was used to improve both the structural fidelity and psychological alignment of generated agents.

Training was performed using techniques such as gradient checkpoint(Feng and Huang, 2021), QLoRA(Dettmers et al., 2023) for parameter-efficient fine-tuning and for model quantization. These methods were applied during both SFT and DPO stages. All experiments were run on Google Colab Pro+ with an A100 GPU.

The summary of the full pipeline (data preparation + model training) it is describe in the Figure below.

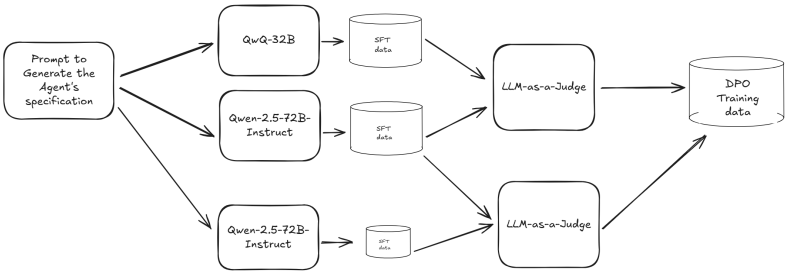


Figure 4: Data Preparation and Model Training Pipeline.

## 5 Results

The evaluation was conducted using a sample of 20 agents selected from the k-means clusters described in the previous section. This sampling strategy was employed to ensure a minimum level of diversity across personality traits.

The evaluation metrics used were as follows:

- **Average Score Alignment:** For each personality trait, the model's responses and the individual's responses to a set of multiple-choice questions were compared. This metric measures the average difference between the model's and the individual's answers, with a lower score indicating greater alignment. A perfect alignment corresponds to a score of zero, and scores can range from 0 to 4. The score alignment for each agent was calculated and then averaged across all agents to obtain this metric.
- **Average Accuracy:** This metric represents the mean accuracy of the agents' responses when compared to the corresponding individuals' records used to generate each agent's specification.
- **Personality Traits Accuracy:** This metric is the average accuracy for each personality trait, aggregated across the 20 test agents.

### 5.1 Quantitative Evaluation

| Model | Average Score Alignment | Average Accuracy |
|---|---|---|
| Trained Qwen | 1.4470 | 19.67% |
| Qwen 7b - Instruct | 1.4688 | 21.46% |
| GPT4O | 1.5006 | 20.00% |

Table 1: Model comparison: Average Score Alignment and Accuracy

### 5.2 Qualitative Analysis

A qualitative examination of the results reveals that, although all models demonstrated reasonable score alignment, none achieved high overall accuracy. Notably, the performance was still inferior to the state-of-the-art method proposed by Zhu et al. (2025).

From the results presented in the table above, it is evident that the trained Qwen model attained the best average score alignment. However, its average accuracy was the lowest among the evaluated models. This indicates that while the Qwen model's answers are, on average, closer to human responses (as measured by the IPIP-120 test), it is less likely to produce exact matches compared to other models.

Combining the quantitative findings from the table with insights from the related heat map, we observe that the Qwen model consistently achieved the highest accuracy for personality traits. Since

the primary goal of this study is to align agent specifications with personality traits, these results suggest that the trained Qwen model is the most effective among those evaluated.

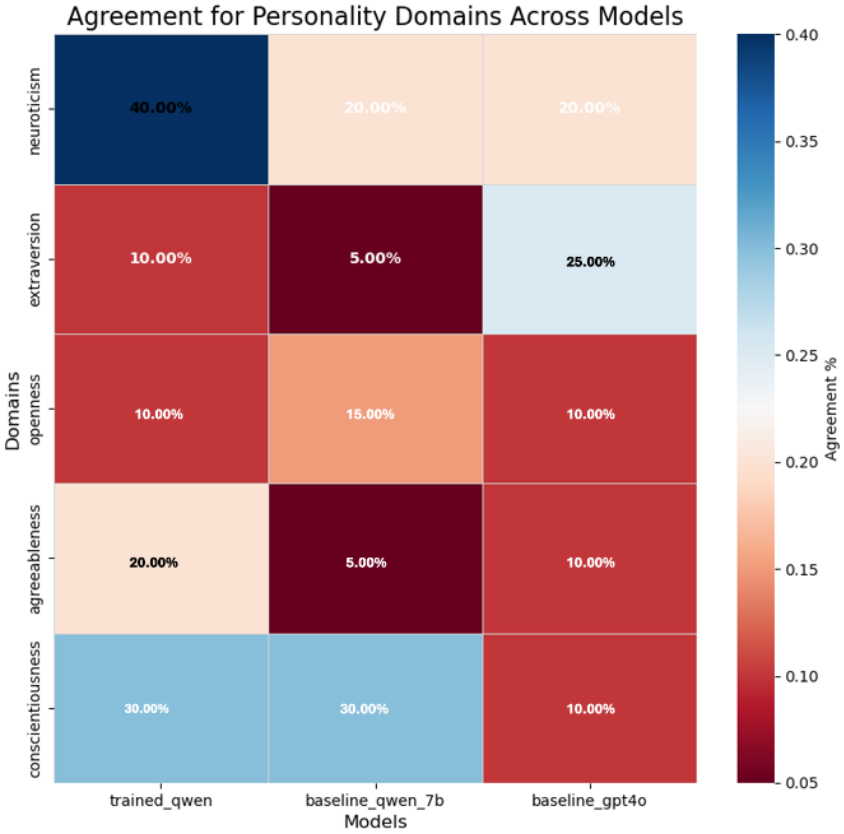


Figure 5: Visualization of results.

## 6 Discussion

Reaching strong, evidence-based conclusions in this study is currently limited by the available data. While the results suggest it might be possible to improve the alignment between agents' descriptions and personality traits, verifying whether this improvement enhances downstream task performance was beyond the scope of this project. For instance, we did not investigate whether agents based on individuals with a high level of conscientiousness perform better than those with a low level of conscientiousness on specific tasks for example.

In comparison to the work of Zhu et al. (2025), which uses prompts instead of persona-based agents, their alignment scores per trait suggest potentially better performance. However, it remains questionable whether these scores translate to an exact match between model and human responses in terms of personality alignment. This raises the broader challenge of whether language models can truly replicate human behavior based solely on personality information. Given the vast amount of data used to train large language models, The model can likely reproduce expected behaviors, but not necessarily the decision-making patterns of a cohort defined by personality traits, sex, gender and country.

Nonetheless, it might be possible to simulate such behaviors by using agents in combination with better agent orchestration, which could help constrain the decision-making workflow of the AI system and lead to better personality alignment. For example, breaking down downstream tasks, formulating prompts to interact with other agents, and employing different decision-making frameworks based on personality traits, agents could potentially enhance performance on downstream tasks.

## 7 Conclusion

In summary, our evaluation across 20 diverse agents demonstrates that the trained Qwen model achieves the highest alignment between agent responses and human personality traits in comparison with the baselines of this project. However, the results highlight that improved alignment does not necessarily translate into exact answer matches.

The available data do not allow for strong conclusions regarding the benefits of model alignment for persona agent-based systems, and further research with larger and more diverse samples is necessary to better understand the practical impact and effectiveness of improving downstream task performance through personality traits alignment.

## 8 Team Contributions

- **Caroline Santos:** This member was responsible and execute all the steps described in this work.

**Changes from Proposal** Originally, the plan was to generate preference pairs by administering the IPIP-120 test to each agent and selecting the one with the most accurate responses as the 'chosen' agent, with the other marked as 'rejected.' However, this approach was not feasible due to time and resource constraints. Running the TinyTroupe framework at scale, particularly for evaluating more than ten agents across multiple iterations is financially expensive.Therefore, the approach LLM-as-judge was used.

## References


Yanquan Chen, Zhen Wu, Junjie Guo, Shujian Huang, and Xinyu Dai. 2024. Extroversion or Introversion? Controlling The Personality of Your Large Language Models. arXiv:2406.04583 [cs.CL] https://arxiv.org/abs/2406.04583

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314 [cs.LG] https://arxiv.org/abs/2305.14314

Jianwei Feng and Dong Huang. 2021. Optimal Gradient Checkpoint Search for Arbitrary Computation Graphs. arXiv:1808.00079 [cs.LG] https://arxiv.org/abs/1808.00079

John A. Johnson. 2014. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality* 51 (2014), 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

Paulo Salem, Christopher Olsen, Paulo Freire, Yi Ding, and Prerit Saxena. 2024. TinyTroupe: LLM-powered multiagent persona simulation for imagination enhancement and business insights. https://github.com/microsoft/tinytroupe. GitHub repository.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Personality Alignment of Large Language Models. arXiv:2408.11779 [cs.CL] https://arxiv.org/abs/2408.11779

Qingmeng Zhu, Tianxing Lan, Xiaoguang Xue, Zhipeng Yu, and Hao He. 2024. TraitsPrompt: Does Personality Traits Influence the Performance of a Large Language Model?. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 2912–2917. https://doi.org/10.1109/CSCWD61410.2024.10580436