

Can Large Language models mimic human personality?

Anonymous ACL submission

Abstract

Personality induction into LLMs aims to simulate human personality traits, with potential applications in personalized interactions and high-quality human-like synthetic data generation. It is thus a promising but challenging frontier in natural language processing. In our study, we use the Essays Dataset (Pennebaker and King, 1999), as its extended narratives are better suited for modeling stable personality traits; shorter texts, by contrast, often reflect mood states (Schwartz et al., 2013) rather than personality. We explore two key aspects: First, we show that different fine-tuning methods significantly reduce the variance observed in psychological test-based evaluations, which in pre-trained models have previously been shown to be unstable (Gupta et al., 2024), thereby making them more reliable. Second, despite this improvement, our results show that personality induction in LLMs suffers from low accuracy when tuned on unguided text, suggesting that such text might lack the nuanced cues essential for an accurate expression of personality. We make the code and dataset openly available for reproducibility¹.

1 Introduction

Personality, as a concept, has long been researched by psychologists for its role in shaping human behavior, emotional expression, and impact on social interactions. Broadly, personality is defined as the consistent patterns of behavior and emotional tendencies that differentiate individuals (Goldberg, 1993; Yarkoni, 2010); personality has been studied using many curated frameworks, the most common of which is the Big Five (McCrae and John, 1992; John et al., 1999). The ability to reliably and reproducibly measure personality has historically relied on

¹https://anonymous.4open.science/r/personality_induction-E288/README.md

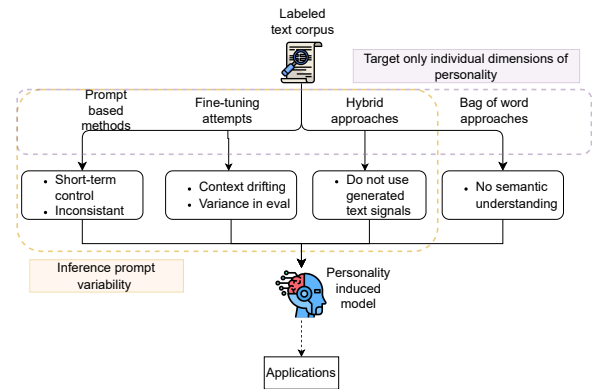


Figure 1: Overview of existing personality induction approaches and their limitations

self-report instruments and behavioral observations by professionals (John et al., 1999).

1.1 The Challenge of Personality Induction in LLMs

With the rise of foundation models, notably large language models (LLMs), a new challenge has emerged: Can these systems be injected with discernible and consistent personality traits such that they can mimic human behavior in their responses? While prior work has largely emphasized personalization (Zhang et al., 2022) and data synthesis (Hämäläinen et al., 2023) as the primary motivations for inducing personality into LLMs, we identify several additional drivers: improving narrative coherence, increasing trust and predictability, enabling controllability in model outputs, deepening human-AI interaction research, and supporting long-term AI identity formation. These motivations underscore the broader utility and urgency of developing principled, stable, and interpretable approaches to personality induction.

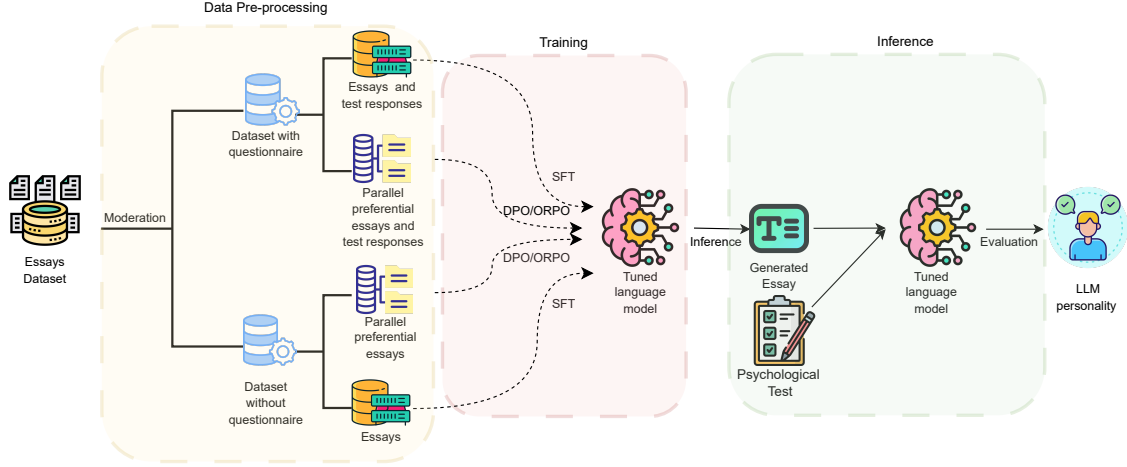


Figure 2: Pipeline for personality induction

1.2 Existing Approaches and Limitations

Efforts to achieve this have resulted in various experimental approaches. In recent times, most of the studies have primarily leveraged controlled prompting techniques to steer LLM outputs for targeted dimensions of personality (Serapio-García et al., 2023; Mao et al., 2023; Caron and Srivastava, 2023; Li et al., 2016). These approaches typically target individual personality dimensions, but their results remain incomplete in capturing the full spectrum of personality traits, and performance in even single dimensions is not particularly impressive. More recent approaches have attempted to induce personality traits implicitly into the LLMs via pretraining and fine-tuning. Some researchers even argue that personality traits can naturally emerge as a secondary result of extensive text-based learning (Hilliard et al., 2024), showing with some models in their work consistent linguistic markers of extraversion, conscientiousness, and other traits. Other researchers argue that the variability in response generation, especially while evaluating the LLMs, raises concerns about the reliability of such evaluation methodologies that are borrowed from psychological literature (Gupta et al., 2024; Frisch and Giulianelli, 2024; Salecha et al., 2024). Figure 1 shows an overview of attempted approaches with their limitations.

Some studies have tried to experiment with whether LLMs can maintain a constant personality profile across varied contexts. For example, studies

in a parallel field, i.e., automated personality recognition, suggest that lingual markers of personality can be quite significantly culture-dependent (Park et al., 2015); the works focused on cross-language analysis to show this (Mairesse et al., 2007; Farnadi et al., 2013). Other researchers have experimented with LLMs to see if they can adapt the style or “persona” considering human feedback (Ouyang et al., 2022). The work showed promise in real-time adjustments (Ziegler et al., 2019). Some others have explored injecting personality via dynamic context adaptation or role-playing, but these efforts also often fizzle out producing inconsistent persona-targeted outputs (Huang et al., 2025; Shanahan et al., 2023). It is important to note that personality evaluation is frequently conducted on individual OCEAN dimensions, each of which inherently possesses a 50% random baseline (Serapio-García et al., 2023; Ouyang et al., 2022). Consequently, reported success metrics may be misleading, as the actual task involves predicting a binary vector across all five OCEAN dimensions, consistent with human evaluation protocols.

1.3 Prior black-box approaches for evaluation and the rationale for using questionnaires for assessment

Using standardized psychological questionnaires closely resembles the approach of professional psychologists, built on decades of validated research, and offers transparency (Podsakoff et al., 2012). Al-

though the questionnaire-based approach has its own downsides such as statistical variability and unclear evidence of its applicability outside of human subjects, we will be using one such test (IPIP-NEO) in our evaluation (Goldberg et al., 1999) (for examples check Table A5 in Appendix) because the accuracies reported on individual traits using simple NLP techniques like TF-IDF and SVMs (Han et al., 2020; Christian et al., 2021) are in range of 60-80% so fitting any sort of classification model can just fit on tokens related to such traits, while psychological tests provide explainability and accuracy results are in the same range (Miotto et al., 2022; Ouyang et al., 2022).

2 Research Questions

- **RQ1: How is the statistical variance in self-test responses affected after fine-tuning?**

Prior studies have highlighted instability in LLM-generated responses to psychometric instruments, with output variability often attributed to prompt sensitivity rather than underlying personality representation (Gupta et al., 2024). This raises concerns about the reliability of current evaluation protocols. Whether fine-tuning reduces response variance is crucial for establishing LLMs as stable subjects in personality assessment.

- **RQ2: Can supervised or preferential fine-tuning using unguided text induce personality in LLMs?**

Much of prior work equates success in personality induction with improvements on individual traits, but whether this reflects coherent, full-spectrum personality remains an open question. We investigate this through fine-tuning approaches and further argue that conditioning on generated essays is essential for trait recovery and capturing nuanced self-expression (Jiang et al., 2023).

- **RQ3: How do personality induction results differ between security-aligned and uncensored models?**

We hypothesized that fine-tuning may influence the model performance of security-aligned models, potentially acting as a confounding variable

in our results. To address this concern, we replicated the methodology on parallel uncensored versions of the models.

3 Dataset and Models used

Personality Trait	True	False
Openness	1,271	1,196
Conscientiousness	1,253	1,214
Extroversion	1,276	1,191
Agreeableness	1,310	1,157
Neuroticism	1,233	1,234

Table 1: Big Five Traits Distribution in the Essays Dataset.

The dataset contains 2,467 essays, totaling 1.9 million words, with an average of approximately 770 words per essay. Samples from the dataset can be seen in section A.7 of the Appendix. We selected this dataset for its rich narratives and longer text lengths, which better capture stable personality traits than shorter sources like Reddit (Gjurković and Šnajder, 2018) or Twitter (Golbeck et al., 2011), which often reflect transient moods (Schwartz et al., 2013).

Model	Size (B)	Input Context (tokens)	Uncensored version used
Gemma-2-2B	2	8,192	Y
Llama3.2-3B	3	128,000	Y
Gemma-7B	7	8,192	N
Llama 3.1-8B	8	8,000	Y
GPT-3.5-turbo-0125	175	16,385	N

Table 2: Comparison of various language models used

We evaluate 5 distinct models, each reflecting different scales and design principles: two small models (LLaMA 3.2–3B (Dubey et al., 2024) and Gemma-2–2B (Team et al., 2024)), two mid-sized models (Gemma–7B (Team et al., 2024) and LLaMA 3.1–8B (Dubey et al., 2024)), and a big commercial model (GPT–3.5²). We compare newer-generation smaller architectures against larger, earlier models to gauge their relative performance in personality induction. We chose these models as they are relatively close in performance for general tasks, as can be seen from the benchmark scores (Dubey et al., 2024; Team et al., 2024).

For RQ3, we compare uncensored versions of these models with their corresponding instruction-tuned counterparts. Specifically, we include the un-

²<https://platform.openai.com/docs/models/gpt-3-5>

censored variants of LLaMA 3.1–8B (Dubey et al., 2024)³, Gemma-2–2B (Team et al., 2024)³, and LLaMA 3.2–3B (Dubey et al., 2024)³. Notably, Gemma-7B currently lacks an uncensored counterpart in the UGI leaderboard³, and is thus excluded from this specific comparison.

4 Methodology

In this study, we adopt a two-phase methodology. We first analyze the statistical variation arising from prompt rephrasing during evaluation of fine-tuned models. Upon establishing prompt stability, we assess the effectiveness of fine-tuning strategies for personality induction. The next subsection outlines our prompt variation protocol, informed by prior work (Gupta et al., 2024), and the second subsection details our fine-tuning methodology and its rationale, plus details about data augmentation, training, and evaluation.

4.1 Assessing the Impact of Prompt Variation

As discussed previously, LLMs can exhibit significant statistical variations when prompted with semantically similar but lexically distinct queries (Gupta et al., 2024). Such variations hinder the assessment of model performance, as it can be unclear if performance changes occur due to genuine learning or prompt-based variabilities. Hence, we took the same prompt templates (shown in Table A2 in the Appendix) as in the previous study and compared the variability between fine-tuned and untuned models; we named them sets S1, S2, and S3 for clarity. We have chosen one of the prompt variations, i.e., S2, to be a string response which is slightly different from previous work (only in one prompt) as it produces more tokens (we hope that this change is not too invasive) and that could be an important variation factor as well because most of the evaluation methods only consider a single token response to calculate the personality score. We use a starter statement, 'My answer to the statement is: ' as it's easy to pick the string for scoring using this template, and in practice, language models are quite consistent in following the template. If the model fails to follow the template and diverges or replies not in the format

being prompted for, we consider that response as "NaN" and it does not count in the evaluation. The NaN rate for all models during evaluation is reported in Table A4 in the Appendix. Results of this experiment are shown in Table 3, and an example of a decrease in variance is shown in Figure 4.

4.2 Training and inference for personality induction

After confirming that the choice of prompt format has negligible impact post fine-tuning, we adopted format S1 for all subsequent experiments. All models underwent two rounds of supervised fine-tuning: (1) using only the prompt and corresponding essays, and (2) incorporating portions of questionnaire items and responses into the prompt to generate essays (see Appendix A.8–A.11 for examples). The rationale for (2) is the hypothesis that the model may implicitly learn associations between essays, questionnaire responses, and personality labels. During evaluation, the fine-tuned model is first prompted to generate an essay, followed by sequential responses to items from a psychological inventory to construct its personality profile. Due to OpenAI’s moderation protocols concerning personal data, approximately 300 samples were filtered out during GPT-3.5 fine-tuning, yielding a final SFT dataset of ≈ 2.1 k samples, used uniformly across all models. Our experiment (see Figure 5 in the Appendix) showed that changing the temperature has a minimal effect on personality accuracy. Based on this, and consistent with prior findings on generation stability (Renze, 2024), we use a temperature of 0.0 during inference, as it is easier to standardize across models. For details regarding hyperparameter choices during training and inference, refer to Appendix Section A.3

4.2.1 Supervised fine-tuning

Each training sample consists of (a) a system prompt specifying what the task is, (b) a user prompt instructing the model to write an essay with certain Big Five personality traits, and (c) the corresponding essay. The model is trained via cross-entropy loss. At inference, it generates an essay in one pass, which is then used as context to predict the corresponding personality label (i.e., the original prompt). Figure 2 illustrates this pipeline, and sample data formats are detailed in Appendix A.9.

³Links: LLaMA 3.1-8B, Gemma-2-2B, LLaMA 3.2-3B, UGI Leaderboard.

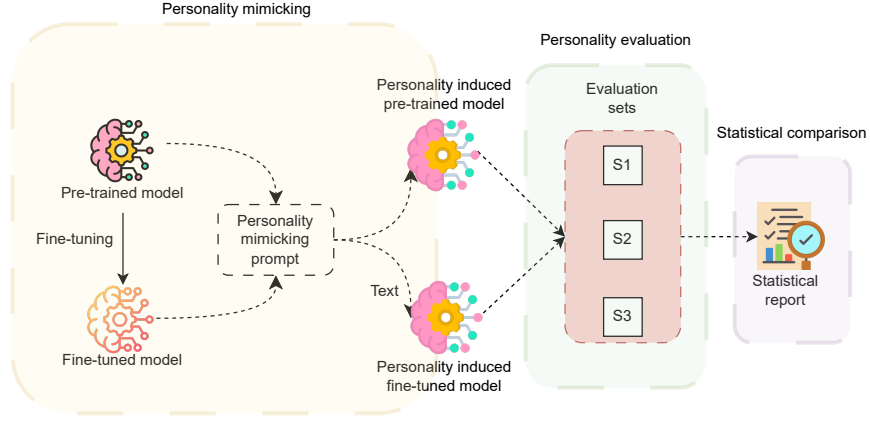


Figure 3: Methodological overview for comparing statistical variation in evaluation questionnaire

4.2.2 Preferential fine-tuning

While SFT aligns the model output to a single “correct” response, preferential fine-tuning includes ranked data that is pairwise assembled to represent human preferences (Ziegler et al., 2019; Rafailov et al., 2024; Hong et al., 2024). This method is similar to Reinforcement Learning with Human Feedback (RLHF) and has proved useful in tasks where the essential difference between “chosen” and “rejected” response is subjective (Ouyang et al., 2022) and thus the model can leverage from these preference signals rather than attempting to fit only one “ground truth” signal. Formally, we implement a preference loss, details of which can be found in the mathematical formulation stated in equations (1), (2), and (3) in the Appendix.

4.2.3 Creating a parallel preferential dataset

Since we have the binary labels for OCEAN personality traits for any given sample, we pick a sample at random such that the OCEAN binary traits are not a perfect match. In our study, we aim to study personality as a whole at the primary target, so nuances in every dimension individually are important. We could pick the exact opposite sample. For example, for a person positive in all the traits, we can pick the preferential parallel data to be negative in all the traits, but in this way, we can never differentiate or teach our model more subtle differences, such as positive in all traits but negative in just one trait. For each given sample, we pick 3 random samples in this way and thus triple the preferential dataset

size ($\approx 6.3k$ parallel essays) for preferential training. Please see Appendix section A.10 for an example.

4.2.4 Using questionnaire while training

We also wanted to test the change in performance after including a part of the evaluation questionnaire itself in training, so that the model learns a link between personality essays and tests. For this task we split the questionnaire into roughly 50% training and test. Then, for each personality essay/type with regards to the OCEAN traits, we generate two sets of question responses: (1) An ideal test for that particular personality type, (2) A randomized test for that particular personality type where the randomization is done such that the average score for OCEAN traits matches the personality. At last, the train set of this questionnaire, alongside their responses, is appended to the input prompt. As is evident, this doubles the dataset size ($\approx 4.2k$ samples), and we then use the same method as explained before to create parallel datasets for preferential training ($\approx 12.6k$ parallel samples). Please see Appendix sections A.8 and A.11 for examples.

5 Results and discussion

5.1 Reduced variance in questionnaire-based evaluations

As shown in Table 3, fine-tuned and preferentially fine-tuned models consistently exhibit lower standard deviation in Big Five personality questionnaire scores compared to their untuned counterparts. This

Versions	Models											
	Llama-3.2-3B			Gemma-2-2B			Gemma-7B			GPT-3.5		
	S1	S2	S3	S1	S2	S3	S1	S2	S3	S1	S2	S3
Pre-trained	1.86	1.65	1.88	1.90	1.91	1.78	1.62	1.68	1.76	1.80	1.52	1.81
SFT (Essays)	1.40	1.41	1.44	1.46	1.44	1.28	1.33	1.38	1.32	1.23	1.24	1.30
SFT (Essays + Q)	1.42	1.30	1.35	1.40	1.28	1.29	1.19	1.29	1.42	1.32	1.22	1.21

Table 3: Variance is calculated over 32 personality types comparing responses of pre-trained vs fine-tuned models. Responses are considered on a scale of 1-5.

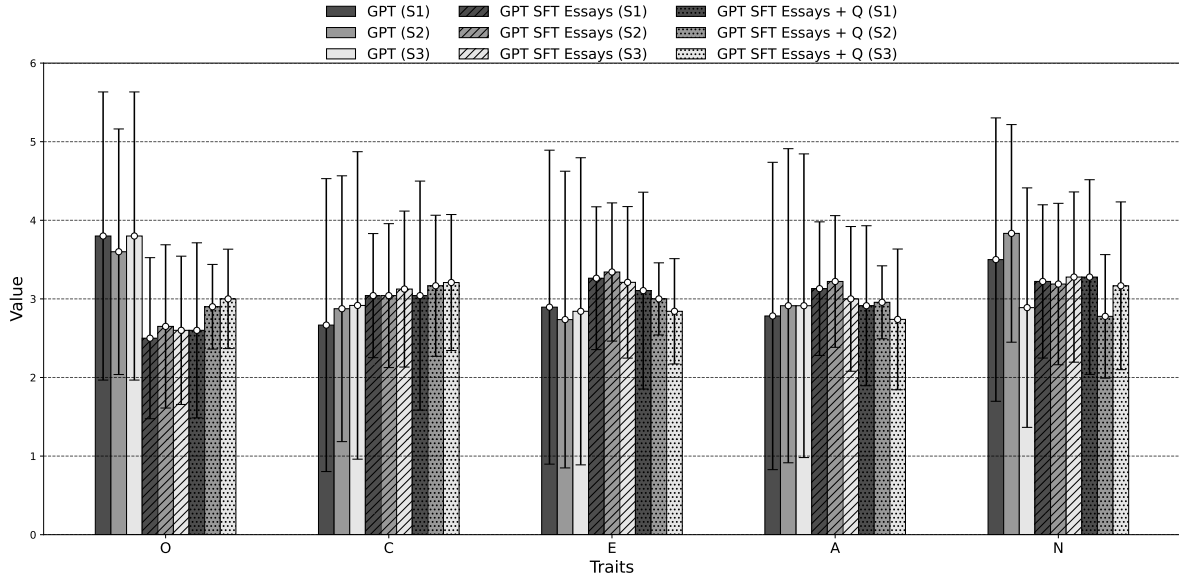


Figure 4: Standard deviation in questionnaire responses for GPT-3.5 across three prompt variations (S1: numeric, S2: string, S3: alphabet) under three settings—base GPT, SFT with personality essays, and SFT with essays plus questionnaire fragments—when conditioned to simulate all positive OCEAN traits

reduction in variability, ranging from approximately 15% to 33% is observed across all models evaluated. We hypothesize that the higher variance reported in prior studies may result from the lack of explicit task-specific supervision during model training (Salecha et al., 2024), with pre-trained models exhibiting heightened sensitivity to prompt phrasing in the absence of such adaptation.

Answer to RQ1

Fine-tuning on labeled human-generated text reduces the variability of LLM response for self-test psychological assessment thus making them more reliable for evaluation.

This result partially addresses the volatile nature of psychological tests because it shows that at least one source of prompt rephrasing based volatility can be mitigated by appropriate training. In a different

environment where such models act as reward models or judging models, this kind of training can be used. Figure 4 is an example that shows the reduction in standard deviation for GPT-3.5 w.r.t changes in the evaluation questionnaire. It is to be noted that our std values are higher than those reported in previous work because we are testing for a particular personality (all positive OCEAN in this example), as this is the fine-tuning task, while they tested the model only for one personality, considering it has to have an inherent personality.

5.2 Low accuracy of post-training methods using unguided text for personality induction

Despite achieving improvements in response stability, fine-tuned models across all methods of post-training fail to show any sort of convincing accuracy

Model	Exact match (S1)				Any 4 dimensions match (S1)			
	Base	SFT	DPO	ORPO	Base	SFT	DPO	ORPO
<i>Without Questionnaires</i>								
gemma-2-2b	3.13% (1)	3.13% (1)	0.00% (0)	3.13% (1)	18.75% (6)	18.75% (6)	12.50% (4)	18.75% (6)
gemma-7b	9.38% (3)	6.25% (2)	3.13% (1)	3.13% (1)	25.00% (8)	21.88% (7)	18.75% (6)	15.63% (5)
llama-3.2-3b	0.00% (0)	3.13% (1)	3.13% (1)	3.13% (1)	12.50% (4)	18.75% (6)	21.88% (7)	18.75% (6)
llama-3.1-8b	6.25% (2)	6.25% (2)	3.13% (1)	6.25% (2)	18.75% (6)	21.88% (7)	18.75% (6)	18.75% (6)
GPT-3.5	3.13% (1)	3.13% (1)	-	-	15.63% (5)	15.63% (5)	-	-
<i>With Questionnaires</i>								
gemma-2-2b	3.13% (1)	3.13% (1)	0.00% (0)	0.00% (0)	18.75% (6)	18.75% (6)	12.50% (4)	12.50% (4)
gemma-7b	9.38% (3)	6.25% (2)	3.13% (1)	3.13% (1)	25.00% (8)	21.88% (7)	18.75% (6)	18.75% (6)
llama-3.2-3b	0.00% (0)	6.25% (2)	3.13% (1)	3.13% (1)	12.50% (4)	21.88% (7)	21.88% (7)	18.75% (6)
llama-3.1-8b	6.25% (2)	3.13% (1)	6.25% (2)	6.25% (2)	18.75% (6)	18.75% (6)	21.88% (7)	18.75% (6)
GPT-3.5	3.13% (1)	6.25% (2)	-	-	15.63% (5)	25.00% (8)	-	-

Table 4: Evaluation results of fine-tuned models with and without questionnaires. Accuracies are shown as percentages while the numbers in brackets are cases out of 32 that passed the self test.

Model	Censored				Uncensored			
	Base	SFT	DPO	ORPO	Base	SFT	DPO	ORPO
<i>Without Questionnaires</i>								
gemma-2-2b	3.13%	3.13%	0.00%	3.13%	0.00%	0.00%	0.00%	3.13%
llama-3.2-3b	0.00%	3.13%	3.13%	3.13%	3.13%	3.13%	3.13%	3.13%
llama-3.1-8b	6.25%	6.25%	3.13%	6.25%	3.13%	9.38%	6.25%	6.25%
<i>With Questionnaires</i>								
gemma-2-2b	3.13%	3.13%	0.00%	0.00%	0.00%	3.13%	3.13%	3.13%
llama-3.2-3b	0.00%	6.25%	3.13%	3.13%	3.13%	6.25%	3.13%	6.25%
llama-3.1-8b	6.25%	3.13%	6.25%	6.25%	9.38%	9.38%	9.38%	9.38%

Table 5: Side-by-side comparison of exact match results (S1) for censored and uncensored models, with and without questionnaires.

in personality induction. Full results can be seen in Table 4. For all methods of training we check the accuracy for all 32 possible cases (2^5 combinations, taking each OCEAN trait as a binary input) and the maximum we’re able to get is 3/32 (9.38%) which is barely an improvement from a random baseline of 3.125% even though on individual traits (see Table A1 in Appendix) we’re able to reproduce metrics close to previous reports(Chen et al., 2024; Miotto et al., 2022; Ouyang et al., 2022). This brings into question the validity of such results, since the task is to induce personality, which is multi-dimensional. Can we regard it as a success to make improvements in singleton dimensions? when even simpler approaches using bag of words can achieve comparable metrics(Han et al., 2020; Christian et al., 2021). Even for a slightly relaxed measure of at-least 4 dimensions matching the ground truth we could only achieve a maximum of 25% accuracy as is shown in Table 4.

Answer to RQ2

Despite extensive experimentation, neither supervised fine-tuning nor preference-based post-training enabled the models to recover the intended Big-Five personality vectors from unguided essays. Across architectures, accuracies hovered near the random baseline, indicating that current methods on unguided text are insufficient for reliable personality induction.

Building on this observation, we contend that prior studies that evaluate personality induction solely at the level of single Big-Five traits risk overstating success. Because the Big-Five construct is defined as a five-dimensional vector, competence on isolated dimensions cannot be assumed to generalise to the joint space where traits interact. Indeed, well-validated psychometric instruments such as the 120-item and 300-item IPIP-NEO reliably measure trait interdependencies and yield highly correlated profiles(r^2

ranging from 0.7 to 0.9) across retest administrations (Maples et al. 2014; TARIGAN et al. 2024). Consequently, an LLM that fails to reproduce the full personality vector lacks face validity when compared with human testing standards.

5.3 Security alignment does not impacts personality induction results

Prior research (Chen et al., 2024; Miotto et al., 2022; Ouyang et al., 2022) has not explicitly examined the influence of security alignment on model performance. However, the results presented in Table 5 indicate that uncensored models exhibit comparable performance to their censored counterparts. This is despite a generally higher incidence of “NaN” outputs, as documented in Table A4 of the Appendix, which may be attributed to a lesser instruction following ability. These findings suggest that security alignment can be reasonably excluded from consideration in performance evaluation when employing post-training techniques on unguided text for personality induction.

Answer to RQ3

Our experiments on the uncensored versions of models show no significant gains over security-aligned models for reliably inducing personality traits. While alignment factors can slightly affect performance in general, for personality induction, they do not appear to be the main bottleneck in boosting accuracy.

6 Conclusion

Our study demonstrates that fine-tuning substantially reduces the variance in self-test questionnaire responses for psychological evaluation of personality-induced LLMs, suggesting that at least part of the volatility reported for pre-trained models can be mitigated after the models learn from supervised signals. Despite improvement in stability, personality induction using unguided text falls short of expectations in terms of accuracy. This points to a need for more targeted datasets that feature human responses in specific scenarios that showcase personality better, or for extended chatbot-style dialogue methods that can accumulate the critical cues needed by prompting the user to provide more data when uncertain about specific dimensions of their personality to make a

more confident assessment. We eliminate the involvement of safety-alignment protocols as a confounding factor in post-training attempts to induce personality. We also conclude from our results that much of the previous work that relies on individual dimensions of personality to support their arguments is not representative of the original task of personality induction, which is multidimensional. Psychological research shows strong adherence of tests like IPIP-NEO to real-world personality, while considering a complete persona profile, and future work with LLMs should also strive to achieve this target or find other ways of reliably evaluating personality in LLMs.

7 Ethics and Responsible Use Statement

This study investigates the induction and evaluation of personality traits in large language models (LLMs), intersecting with critical ethical domains such as human data privacy, psychological well-being, and responsible AI deployment. We affirm that all models and datasets employed in this work were used exclusively for academic research and educational purposes. No commercial use whether of model weights, outputs, or derivatives is pursued or endorsed.

Data and Licensing Compliance

The dataset utilized in this work, the Essays Dataset (Pennebaker and King, 1999), is publicly available and used in accordance with its licensing terms (Apache License 2.0), which permit redistribution and modification for research purposes. The personality trait labels and fine-tuned model variants derived from this dataset are strictly intended for methodological evaluation and not for clinical or diagnostic purposes.

All model resources were employed under appropriate non-commercial licenses:

- **Gemma Models:** Distributed under the Gemma Community License, which restricts usage to non-commercial research and prohibits production-level deployment.
- **LLaMA Models:** Provided under the Meta LLaMA 3 Community License, allowing access to academic and research institutions only.

- **GPT-3.5 (OpenAI):** Accessed via API under OpenAI’s Usage Guidelines, allowing limited research use while prohibiting training or redistribution of outputs for commercial purposes.

All training and evaluation data were sourced from publicly available repositories or licensed datasets with appropriate authorization. No private, proprietary, or personally identifiable information (PII) was included. Our work is compliant with the EU General Data Protection Regulation (GDPR) and prevailing ethical norms.

Ethical Considerations and Oversight

We recognize the dual-use nature of personality modeling. Such techniques may be misappropriated for impersonation or social engineering. As a precaution, we strongly discourage any deceptive or manipulative application of personality-simulating models and urge the broader community to develop and adopt transparent safety mechanisms. Public disclosure of model usage contexts and safeguards is essential to ensure responsible deployment.

The open-source release of all developed resources—including data processing pipelines, prompts, evaluation code, and bias analysis methodology—is licensed under the MIT License. Documentation will accompany the release to promote reproducibility and transparency.

Finally, we commit to ongoing ethical oversight, including periodic reevaluation of datasets, generation behaviors, and potential risks. Our efforts are aligned with environmental sustainability objectives, including the reduction of computing-related emissions in accordance with carbon neutrality goals.

8 Limitations

- **Small, Coarsely Labeled Dataset:** The Essays Dataset (Pennebaker and King, 1999), with 2,467 essays totaling 1.9 million words (average 770 words per essay), is larger and more narrative-rich than many datasets used in prior personality induction studies, such as Reddit posts (Gjurković and Šnajder, 2018) or Twitter/X data (Golbeck et al., 2011), which often consist of thousands of short texts (e.g., 10,000–50,000 samples, typically under 280 characters). These smaller texts can lead to

models learning superficial keyword patterns, as simple bag-of-words approaches like TF-IDF with SVMs already achieve comparable performance (60–80% accuracy on individual traits) (Han et al., 2020; Christian et al., 2021). The Essays Dataset’s longer, introspective narratives prioritize quality, capturing deeper personality signals, but its modest size may still limit linguistic diversity, and its binary Big Five labels may oversimplify trait gradients (Goldberg, 1993). Moreover, high-quality, large-scale datasets with rich personality annotations are scarce, and generating synthetic data often oversimplifies the task by lacking the nuanced, context-dependent expressions found in human-generated text (Hämäläinen et al., 2023; Miotto et al., 2022). As mentioned earlier, if we mix smaller datasets consisting of tweets, Reddit posts, etc, into our work, it has a thread of polluting the results without achieving the original task of capturing personality, as short texts tend to be more "mood" driven.

- **Questionnaire Based Evaluation Bias:** While being grounded in psychology literature, Big Five questionnaires were intended for humans to self-report, and thus the low accuracy might reflect an inherent mismatch between introspection capabilities in humans vs. probabilistically generated tokens (Podsakoff et al., 2012; Gow et al., 2005).
- **No Human Evaluation:** Our study did not include cross-verification of the generated text by human experts, leaving the subjective quality and authenticity of personality expression in the outputs unassessed. Human evaluation is critical for determining whether the model’s generated text convincingly reflects targeted personality traits, such as linguistic markers of extraversion or conscientiousness, and passes an "eye check" for human-like quality. Research indicates that human evaluators often exhibit bias against AI-generated text when aware of its origin, rating it less favorably compared to human-authored content (Christian et al., 2021). Despite this potential bias, human validation would have provided valuable insights into

whether our model’s outputs align with psychological expectations of personality expression. Given the negative results of our experiments, which showed limited success in inducing consistent personality traits, and the economic burden of recruiting expert evaluators, we refrained from conducting this resource-intensive task.

- **Threat of Overfitting** The small scale of the Esays Dataset and the multiplicative use of data in post-preferential fine-tuning increase the risk of overfitting, where the model may memorize patterns rather than generalize personality traits.
- **Non-extensive hyperparameter tuning and single training runs:** Given the extensive duration required for each individual experiment (approximately 24 hours) and the substantial number of experiments conducted, we were unable to perform multiple training runs with systematically varied hyperparameters. This constraint limited our capacity to thoroughly evaluate the sensitivity of our models to different training configurations.

References

- Graham Caron and Shashank Srivastava. 2023. [Manipulating the perceived personality traits of language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2370–2386, Singapore. Association for Computational Linguistics.
- Yanquan Chen, Zhen Wu, Junjie Guo, Shujian Huang, and Xinyu Dai. 2024. Extroversion or introversion? controlling the personality of your large language models. *arXiv preprint arXiv:2406.04583*.
- Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Z Zamli. 2021. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8(1):68.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. 2013. How well do your facebook status updates express your personality. In *Proceedings of the 22nd edition of the annual Belgian-Dutch conference on machine learning (BENELEARN)*, page 88.
- 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*.
- Matej Gjurković and Jan Šnajder. 2018. 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*, pages 87–97.
- Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting personality from twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 149–156. IEEE.
- Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.
- Lewis R Goldberg et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.
- Alan J Gow, Martha C Whiteman, Alison Pattie, and Ian J Deary. 2005. Goldberg’s ‘ipip’ big-five factor markers: Internal consistency and concurrent validation in scotland. *Personality and Individual Differences*, 39(2):317–329.
- Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of llm personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314.
- Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.
- Songqiao Han, Hailiang Huang, and Yuqing Tang. 2020. Knowledge of words: An interpretable approach for personality recognition from social media. *Knowledge-Based Systems*, 194:105550.
- Airlie Hilliard, Cristian Munoz, Zekun Wu, and Adriano Soares Koshiyama. 2024. Eliciting big five personality traits in large language models: A textual analysis with classifier-driven approach. *arXiv preprint arXiv:2402.08341*.

- Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*.
- Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.
- Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.
- François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.
- Shengyu Mao, Ningyu Zhang, Xiaohan Wang, Mengru Wang, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2023. Editing personality for llms. *arXiv preprint arXiv:2310.02168*.
- Jessica L Maples, Li Guan, Nathan T Carter, and Joshua D Miller. 2014. A test of the international personality item pool representation of the revised neo personality inventory and development of a 120-item ipip-based measure of the five-factor model. *Psychological assessment*, 26(4):1070.
- Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.
- Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Gregory Park, H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Michal Kosinski, David J Stillwell, Lyle H Ungar, and Martin EP Seligman. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6):934.
- James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.
- Philip M Podsakoff, Scott B MacKenzie, and Nathan P Podsakoff. 2012. Sources of method bias in social science research and recommendations on how to control it. *Annual review of psychology*, 63(1):539–569.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356.
- Aadesh Salecha, Molly E Ireland, Shashanka Subrahmanya, João Sedoc, Lyle H Ungar, and Johannes C Eichstaedt. 2024. Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.
- H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.
- Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

MEDIANITA TARIGAN et al. 2024. The validity and reliability of the Indonesian IPIP-NEO-120. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*, 31(1).

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.

Shiwei Zhang, Yiyang Du, Guanzhong Liu, Zhao Yan, and Yunbo Cao. 2022. **G4: Grounding-guided goal-oriented dialogues generation with multiple documents**. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 108–114, Dublin, Ireland. Association for Computational Linguistics.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Appendix

A.1 Evaluation on individual dimensions

While our primary focus is on multi-dimensional personality induction, we also evaluate model performance on each individual dimension of the OCEAN framework—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism—to facilitate comparison with prior work.

As shown in Table A1 in the Appendix, individual trait prediction results vary considerably across models and training strategies. The highest accuracies are observed in the range of 65%. These results are broadly consistent with findings in prior literature (Chen et al., 2024; Miotto et al., 2022; Ouyang et al., 2022), though our peak scores are slightly lower. One plausible explanation for this discrepancy is the reduced statistical power in our evaluation setup, we utilize only 32 test instances per trait, compared to more expansive test sets used in previous studies.

Despite the relatively modest accuracy scores, performance is consistently above the random baseline of 50% for binary classification, which reaffirms that

certain linguistic markers of personality are retained in fine-tuned models. However, these results must be interpreted cautiously. Improvements in isolated dimensions do not necessarily indicate meaningful success in personality induction unless the joint multi-dimensional structure of personality is preserved. As elaborated in the main paper, our full-vector evaluations reveal that models fail to jointly reproduce personality profiles with statistical reliability, and thus, trait-level gains must not be overstated.

A.2 Prompt sets for self-test evaluation

Table A2 in Appendix comprehensively shows the sets S1, S2, and S3 used in our study for comparison of statistical variations while evaluating.

A.3 Training Steps and Hyperparameters

All experiments were conducted using an NVIDIA A100 80GB GPU under PyTorch 2.0. We employed the trl library for both supervised fine-tuning (SFT) and preference-based methods (DPO, ORPO).

Additional environment details for reproducibility:

- **Python Version:** 3.10

- **CUDA Version:** 12.1

- **Transformers Library Version:** 4.40.2

- **Operating System:** Ubuntu 22.04

Supervised Fine-tuning (SFT). We used standard LoRA (Hu et al., 2022) for SFT. Across 3 epochs:

- **Learning Rate:** 1×10^{-5}

- **Batch Size:** 2 (with gradient accumulation step of 4)

- **LoRA Configuration:** Rank (r) of 8, dropout = 0.1

- **Precision:** 16-bit (FP16)

- **Max Prompt/Sequence Length:** 4,000 tokens

- **Optimizer:** AdamW ($weight\ decay = 0.01$, $betas = (0.9, 0.999)$)

- **Scheduler:** Cosine scheduler with linear warm-up ($warmup\ ratio = 0.1$)

Preference-Based Fine-tuning (DPO, ORPO).

For both DPO and ORPO, we relied on QLoRA(Dettmers et al., 2023) to reduce memory usage. We trained for 3 epochs with:

- **Learning Rate:** 1×10^{-5}
- **Batch Size:** 2 (with gradient accumulation steps of 4)
- **QLoRA Configuration:** 4-bit quantization (*nf4*), rank (*r*) of 8, dropout = 0.1
- **Max Prompt/Sequence Length:** 4,000 tokens
- **Optimizer:** AdamW (*weight decay* = 0.01, *betas* = (0.9, 0.999))
- **Scheduler:** Cosine scheduler with linear warm-up (*warmup ratio* = 0.1)

Inference Settings. For all models and methods, we set:

- **Max Length:** 200
- **Sampling:** *temperature* = 0.0, *top_p* = 1.0, *do_sample* = False

The inference settings are common for all models; a temperature of 0.0 and a top-p of 1.0 are used. We chose a temperature of 0.0 because it minimizes generation variability and makes evaluation more consistent across runs. There is not much variability with a change in temperature. For example, have a look at Figure 5, the percent change in accuracy is in the domain of 0-6%. This is consistent with other reported results(Renze, 2024), thus we use this temperature. With temperature set to 0.0, the nucleus sampling parameter top-p becomes inconsequential, as sampling is effectively disabled, ensuring a controlled and reproducible inference environment across all models.

Typically, small-scale models require approximately 14–16 hours for a single round of supervised fine-tuning (SFT) and 22–24 hours for a single round of ORPO training. In comparison, mid-sized models take approximately 20–22 hours for SFT and 26–32 hours for ORPO. Given the substantial number of experiments conducted and the associated computational demands, we executed each training configuration only once, with the exception of some inference runs as reported in Figure 5.

A.4 Negative-Phrased Questions and Score Inversion

In our questionnaire, we deliberately include negatively phrased items to mitigate simple pattern matching. For example, an Extravert might naturally agree (i.e., choose a high value) with a statement like “*I enjoy meeting new people*”. However, for a negatively phrased counterpart such as “*I am not good at getting people to like me*,” the expected response is a low value (indicating disagreement).

During scoring, these negatively phrased items are inverted so that a low response (e.g., 1) is automatically mapped to a high score (e.g., 5), ensuring the correct trait direction. This helps detect whether models are genuinely interpreting question semantics rather than relying on simple keyword matching or always selecting high values. The questions include placeholders (math field) to indicate such reversals, where negative statements are inverted by applying ($6 - response$) in inference when math value = -1 (check Table A5 in Appendix for samples).

A.5 Reflecting human preferences via preferential fine-tuning

Preferential fine-tuning leverages the idea that human judgments about preferred outputs not only provide label quality but also embed nuanced, context-specific values. By assigning higher weights to chosen responses and lower ones to rejected options, the model internalizes preferences that often mirror an individual’s or organization’s stylistic or ethical priorities. This calibration is particularly useful for tasks like personality induction, where alignment with specific traits or tones is key: user preferences effectively “shape” the model’s persona. We chose this approach as it respects the principle of human-centric design—amplifying desired personality cues grants practitioners a more direct avenue to impart generative models with human-like qualities that reflect real-world norms and relationships in language. The two methods picked, i.e., DPO and ORPO, are explained below.

A.5.1 Direct Preference Optimization (DPO)

DPO is one of the widely used methods (Rafailov et al., 2024) where the model itself also acts as a reward model on parallel preferential data. The model compares two candidate outputs, i.e., “cho-

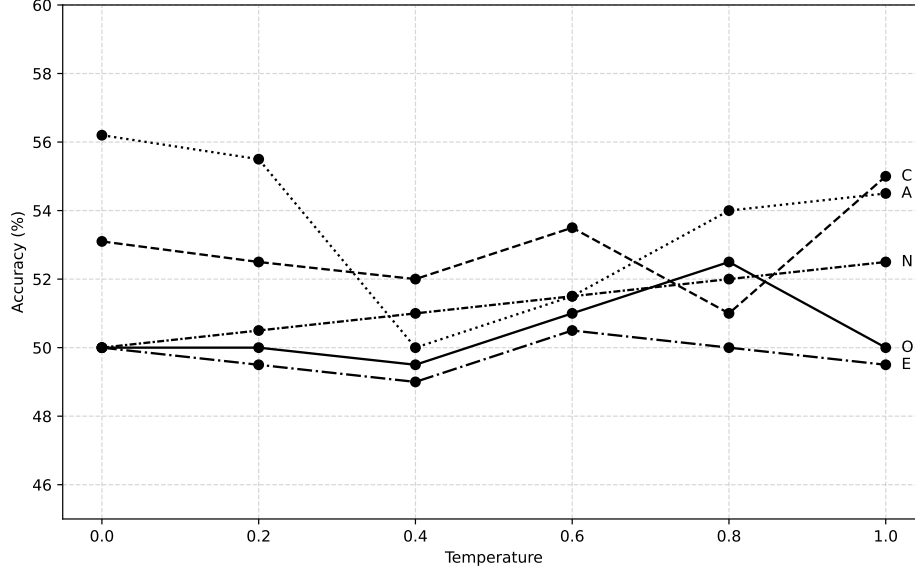


Figure 5: GPT-3.5 SFT OCEAN Trait Accuracy vs Temperature

$$L_{\text{DPO}}(\theta) = -\mathbb{E}_{(x, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \left([\log \pi_{\theta}(y_w | x) - \log \pi_{\theta}(y_l | x)] - [\log \pi_{\text{ref}}(y_w | x) - \log \pi_{\text{ref}}(y_l | x)] \right) \right) \right] \quad (1)$$

$$L_{\text{ORPO}}(\theta) = \mathbb{E}_{(x, y_w, y_l) \sim D} [L_{\text{SFT}}(\theta; x, y_w) + \lambda \cdot L_{\text{OR}}(\theta; x, y_w, y_l)] \quad (2)$$

$$L_{\text{OR}}(\theta; x, y_w, y_l) = -\log \sigma(\log \pi_{\theta}(y_w | x) - \log \pi_{\theta}(y_l | x)) \quad (3)$$

sen" vs "rejected"; it uses a log-likelihood ratio that assigns a higher probability to preferred outputs. Equation 1 shows the DPO loss function where: $L_{\text{DPO}}(\theta)$: The loss function with respect to the policy parameters θ . $\mathbb{E}_{(x, y_w, y_l) \sim D}$: Expectation over the dataset D containing prompts x and pairs of responses (y_w, y_l) where y_w is the preferred response over y_l . $\sigma(z) = \frac{1}{1 + e^{-z}}$: The sigmoid function. β : A scaling factor (inverse temperature) controlling the strength of the update. $\pi_{\theta}(y | x)$: The probability assigned by the policy π_{θ} to response y given prompt x . $\pi_{\text{ref}}(y | x)$: The probability assigned by the reference policy π_{ref} to response y given prompt x .

A.5.2 Monolithic Preference Optimization (ORPO)

ORPO (Hong et al., 2024) targets to achieve the same task as DPO, but it uses an odds ratio of probabilities for token prediction, which is done "offline,"

meaning the gradient is calculated normally, unlike DPO, where it is included in the log-likelihood calculations while backpropagation, instead in ORPO the model used the odds ratio to provide weights for chosen and rejected responses. Equations 2 and 3 show the ORPO loss function where: $L_{\text{ORPO}}(\theta)$ is the overall loss function with respect to model parameters θ . $\mathbb{E}_{(x, y_w, y_l) \sim D}$ denotes the expectation over the dataset D containing prompts x and pairs of responses (y_w, y_l) , where y_w is the preferred response and y_l is the less preferred response. $L_{\text{SFT}}(\theta; x, y_w)$ represents the supervised fine-tuning loss for the preferred response y_w given prompt x . λ is a hyperparameter controlling the balance between L_{SFT} and L_{OR} . $L_{\text{OR}}(\theta; x, y_w, y_l)$ is the relative ratio loss. $\sigma(z) = \frac{1}{1 + e^{-z}}$ is the sigmoid function. $\pi_{\theta}(y | x)$ is the probability assigned by the model θ to response y given prompt x .

A.6 Looking closely at evaluation

Here is an example of Llama-3.2-3B model tuned with ORPO (Essays + Q) and being evaluated for all positive OCEAN traits. Looking at the response to the statement "Am annoyed by others' mistakes" we can notice that the justification (we specifically prompted for it) is valid, but it contradicts the response to statements "Am inclined to forgive others." and "Am not disturbed by events." which are as expected as the model is prompted to be positive in Agreeableness. This is typical, though, that the model is justifying its answer correctly but swinging in its internalization of how an agreeable person should behave. On other occasions, though, it can give a neutral response, too, as to the statement "Am easily offended." which is not a problem as the model indeed should learn to be uncertain on certain topics and not overfit or "hack" rewards from similar questions seen during training.

A.7 Essays Dataset Samples

1997_504851.txt

Traits: EXT: y NEU: y AGR: y CON: n
OPN: y

Essay:

Well, right now I just woke up from a mid-day nap. It's sort of weird, but ever since I moved to Texas, I have had problems concentrating on things. I remember starting my homework in 10th grade as soon as the clock struck 4 and not stopping until it was done. Of course it was easier, but I still did it. But when I moved here, the homework got a little more challenging and there was a lot more busy work, and so I decided not to spend hours doing it, and just getting by. But the thing was that I always paid attention in class and just plain out knew the stuff, and now that I look back, if I had really worked hard and stayed on track the last two years without getting lazy, I would have been a genius, but hey, that's all good. It's too late to correct the past, but I don't really know how to stay focused in the future. The one thing I know is that when people say that b/c they live on campus they can't concentrate, it's b. s. For me it would be easier there, but alas, I'm living at home under the watchful eye of my parents and a little nagging sister that just nags and nags and nags. You get my point. Another thing is, is that it's just

a hassle to have to go all the way back to school to just to go to library to study. I need to move out, but I don't know how to tell them. Don't get me wrong, I see where they're coming from and why they don't want me to move out, but I need to get away and be on my own. They've sheltered me so much and I don't have a worry in the world. The only thing that they ask me to do is keep my room clean and help out with the business once in a while, but I can't even do that. But I need to. But I got enough money from UT to live at a dorm or apartment next semester and I think I'll take advantage of that. But off that topic now, I went to sixth street last night and had a blast. I haven't been there in so long. Now I know why I love Austin so much. When I lived in VA, I used to go up to DC all the time and had a blast, but here, there are so many students running around at night. I just want to have some fun and I know that I am responsible enough to be able to have fun, but keep my priorities straight. Living at home, I can't go out at all without them asking where? with who? why? when are you coming back? and all those questions. I just wish I could be treated like a responsible person for once, but my sister screwed that up for me. She went crazy the second she moved into college and messed up her whole college career by partying too much. And that's the ultimate reason that they don't want me to go and have fun. But I'm not little anymore, and they need to let me go and explore the world, but I'm Indian; with Indian culture, with Indian values. They go against "having fun." I mean in the sense of meeting people or going out with people or partying or just plain having fun. My school is difficult already, but somehow I think that having more freedom will put more pressure on me to do better in school b/c that's what my parents and ultimately I expect of myself. Well it's been fun writing, I don't know if you go anything out of this writing, but it helped me get some of my thoughts into order. So I hope you had fun reading it and good luck TA's.

1997_605191.txt

Traits: EXT: n NEU: n AGR: y CON: n
OPN: n

Essay: Well, here we go with the stream of consciousness essay. I used to do things like this in high school sometimes. They were pretty interesting, but

1075	I often find myself with a lack of things to say. I	I'm out of shape. I'd like to get in better shape, but	1123
1076	normally consider myself someone who gets straight	I hate running. It's too dull for me. Hmmm. it's	1124
1077	to the point. I wonder if I should hit enter any time	almost over now. Just a few more minutes. Let's	1125
1078	to send this back to the front. Maybe I'll fix it later.	see if I make it to the next line. Short reachable	1126
1079	My friend is playing guitar in my room now. Sort	goals! Whatever. Anyway, what else do I have to do	1127
1080	of playing anyway. More like messing with it. He's	tonight. I guess I could read some. My shirt smells	1128
1081	still learning. There's a drawing on the wall next to	like dinner. It's pretty disgusting. I need to wake	1129
1082	me. Comic book characters I think, but I'm not sure	up for a 9:30 am class tomorrow. I remember when	1130
1083	who they are. It's been a while since I've kept up	that wasn't early at all. Well, I made it to the next	1131
1084	with comic's. I just heard a sound from ICQ. That's a	line. I'm so proud of myself. That's sarcasm, by the	1132
1085	chat program on the internet. I don't know too much	way. I wonder if I was suppose to right this thing as	1133
1086	about it so I can't really explain too well. Anyway, I	a narrative. Oh well too late now. Time for me to	1134
1087	hope I'm done with this by the time another friend	head out. Until next time, good bye and good luck. I	1135
1088	comes over. It will be nice to talk to her again. She	don't know.	1136
1089	went home this weekend for Labor Day. So did my		
1090	brother. I didn't go. I'm not sure why. No reason	1997_687252.txt	1137
1091	to go, I guess. Hmm. when did I start this. Wow,	Traits: EXT: n NEU: y AGR: n CON: y	1138
1092	that was a long line. I guess I won't change it later.	OPN: y	1139
1093	Okay, I'm running out of things to talk about. I've	Essay: An open keyboard and buttons to push.	1140
1094	found that happens to me a lot in conversation. Not a	The thing finally worked and I need not use periods,	1141
1095	very interesting person, I guess. Well, I don't know.	commas and all those thinks. Double space after a	1142
1096	It's something I'm working on. I'm in a class now	period. We can't help it. I put spaces between my	1143
1097	that might help. The phone just rang. Should I get	words and I do my happy little assignment of jibber-	1144
1098	it? The guy playing the guitar answered it for me.	jabber. Babble babble babble for 20 relaxing minutes	1145
1099	It's for my roommate. My suitemate just came in	and I feel silly and grammatically incorrect. I am	1146
1100	and started reading this. I'm uncomfortable with	linked to an unknown reader. A graduate student	1147
1101	that. He's in the bathroom now. You know, this is	with an absurd job. I type. I jabber and I think about	1148
1102	a really boring piece of literature. I never realized	dinoflagellates. About sunflower crosses and about	1149
1103	how dull most everyday thoughts are. Then again,	the fiberglass that has be added to my lips via clove	1150
1104	when you keep your mind constantly moving like	cigarettes and I think about things that I shouldn't	1151
1105	this, there isn't really time to stop and think deeply	be thinking. I know I shouldn't be thinking. or	1152
1106	about things. I wonder how long this is going to be.	writing let's say/ So I don't. Thoughts don't solidify.	1153
1107	I think it's been about ten minutes now. Only my	They lodge in the back. behind my tongue maybe.	1154
1108	second line. How sad. Well, not really considering	Somewhere at the point of hiding but dinoflagelates,	1155
1109	how long these lines are. Anyway, I wonder what	protistas and what was that sea weed. I think about	1156
1110	I'm going to do the rest of the night. I guess there's	the San Luiz valley and I think about the mushrooms	1157
1111	always homework to do. I guess we'll see. This seat	in cow shit. I think about the ticos and I think about	1158
1112	is uncomfortable. My back sort of hurts. I think I'm	the chiggers that are living in my legs. I itch. I coat	1159
1113	going to have arthritis when I get older. I always	myself with clear nail polish in hopes to suffocate	1160
1114	thought that I wouldn't like to grow old. Not too	the bugs that are living in my legs and I remember	1161
1115	old, I suppose. I've always been a very active person.	Marco. I remember Ecuador and I think about my	1162
1116	I have a fear of growing old, I think. I guess it'll	thoughts and what I am not supposed to be doing in	1163
1117	go away as I age gradually. I don't know how well	this assignment. Thoughts. I wonder if I think in	1164
1118	I'd deal with paralysis from an accident though. As	sentences I wonder what affect my slowish typing	1165
1119	long as I have God and my friends around, I'll be	has on my stream of consciousness and I wonder if	1166
1120	okay though. I'm pretty thirsty right now. There isn't	there is a way that typing speed can be measured	1167
1121	much to drink around my room. Ultimate Frisbee, I	in this study so that so link some generalization of	1168
1122	haven't played that all summer. Fun game, but tiring.	dorky 301 psyc students. green and the table in my	1169

1170	kitchen makes me want to vomit. orange. What an	1997_568848.txt	1218
1171	absurd color. wish I wasn't in the united state. My		
1172	greencard runs out in a few years wonder what I	Traits: EXT: y NEU: n AGR: y CON: y	1219
1173	do. I hope Dr. Linder gets back in his lab because	OPN: n	1220
1174	I really need to find out if he has funds to pay me.	Essay: I can't believe it! It's really happening!	1221
1175	May have to go back to the library. Brainless job	My pulse is racing like mad. So this is what it's	1222
1176	of nothingness that would make me wallow in the	like. now I finally know what it feels like. just a	1223
1177	world of boredom which isn't entirely bad. Need	few more steps. I wonder if he is going to get any	1224
1178	to focus on school organics and such. Period. Two	sleep tonight!? I sure won't! Well, of course I have	1225
1179	spaces after the period. Mistakes and I want to eat	a million deadlines to meet tomorrow so I'll be up	1226
1180	not hungry and I wonder how many people talk about	late anyway. But OH! I'm so so excited! Yes! Yes!	1227
1181	food in there little computer ramblings Feel open and	I can't believe it is finally happening. Wait! Calm	1228
1182	Happy that I am not having to edit this. Type type I	down. We aren't officially a couple yet. What if I end	1229
1183	don't know what I am think Hannah Imi and Osdprey	up not liking him? That would be horrible. Oh great,	1230
1184	house. I remember when I went down to that . she	I wonder how long it'll take me to finish those Cal-	1231
1185	had spiders on hurt wall pain all over the place and	culus problems? I'll get it done. Don't you always,	1232
1186	we painted clouds on the ceiling and the blue walls	Amy? I can't believe Bob did it! He really did it! He	1233
1187	were so obnoxious. Carey. Sex sex sex. yeah. This	is THE miracle worker. If things turn out all right I	1234
1188	is a strange assignment and Portonoy's complaint	will owe him more than I can ever repay. I wonder	1235
1189	is ringing in my head. Eager to finish so that I can	what Steve is doing in Malaysia right now? An entire	1236
1190	start for Whom the Bell Tolls and get on with it.	month! I'll likely clean out his refrigerator by then.	1237
1191	Bio and Carbon atoms bonds and orbitals. Thinking	Omigosh! Food, lunch tomorrow, what will I ever	1238
1192	about the electron configuration that surrounds the	say to him? He is perfect in every way imaginable. It	1239
1193	last letter in my first name and I think that I must have	is so important for him to think of me the same way.	1240
1194	been granted a full "s" orbital one up and one down.	well, maybe not Perfect, but certainly dynamic. Who	1241
1195	spinning on opposite directions and I am thinking	would have ever thought! Good things do indeed	1242
1196	about Scottish poetry about Mike in his kilt and about	come to those who wait! Oh, I'll have to remember	1243
1197	my guitar that I am slowly slowly slowly learning to	to sign the poster he made tomorrow morning. I hope	1244
1198	play. I wonder what goes on in this study. I wonder	Steve's alarm clock is reliable and I don't oversleep.	1245
1199	if those happy little bored entertained grad students	That would be tragic if I slept 'til noon and missed	1246
1200	will scan words and I wonder how I can mess up	the lunch. Thank goodness Portia is coming along. I	1247
1201	this study? Random words like . don't know. ;Me	will definitely need her support as well as Bob's. just	1248
1202	me me me me and I wish that some things were	having her there will take away some of the tension	1249
1203	easier and I wish that I had been keeping my eye on	and put me more at ease. I'll have to rehearse what	1250
1204	the clock. Wondering how long I have been typing	I say beforehand. things can only get better from	1251
1205	and wishing that I was finished because I need to	here, right? hopefully. oh, I'm so nervous! He will	1252
1206	find out if I have to / will work in the Botany lab	be too. maybe even more so. it'll be ok. Why in	1253
1207	again and all that . ILS Belly and the Flamenco.	the world do humans put themselves through such	1254
1208	Bjork and Rozamond Cockrill kickin' it in Saratoga	torture. maybe love is really worth it?	1255
1209	Springs. I hate Molly's cat and wish that it could be		
1210	exchanged for a worthwhile ferret. Type type type.	1997_688160.txt	1256
1211	I have managed to waste over 20 minutes of time I		
1212	think. Who knows. What If I was to write this out	Traits: EXT: y NEU: n AGR: y CON: n	1257
1213	and it took 30 minutes to write and 15 minutes to	OPN: y	1258
1214	type. Thinking about nothing and wishing that some	Essay: Well, here I go with the good old stream	1259
1215	financial aid would come my way. Need a job and a	of consciousness assignment again. I feel like I'm	1260
1216	sprinkling of time. Time to go and sign outta here.	back in freshman HS English class again. Not that	1261
1217	trees	that's a bad thing, mind you, but my English teacher	1262
		freshman year made us do these assignments con-	1263

1264	stantly, and mine were always completely ridiculous,	1312
1265	like, "wow, I'm really hungry. I wish I could go to	1313
1266	Taco Bell. " They really had no point, except as busy	1314
1267	work. In a psychology class, though, I can see the	1315
1268	reasoning behind an assignment like this. Just letting	1316
1269	my mind go free, and putting my random thoughts	1317
1270	down in writing could be a big help in figuring out	1318
1271	why I'm such a psychological screw-up. Well, that's	1319
1272	not true. I don't want y'all getting the wrong idea	1320
1273	about me, being that today was the first day of class	1321
1274	and all. I'm really not a nut case. People may think	1322
1275	I am, but really, I'm a normal kind of gal. Actually,	1323
1276	down here in Texas, I guess I'm not normal. I don't	1324
1277	like to eat biscuits and gravy for breakfast, and coun-	1325
1278	try fried steak with fried okra for dinner. I'm from	1326
1279	Connecticut, and we don't even HAVE okra, much	1327
1280	less worship it like it's some kind of vegetable god-	1328
1281	dess. My mind is starting to go blank—performance	1329
1282	pressure I guess. I'm on the spot here—I don't want	1330
1283	you all to be bored while you're reading this, if you	1331
1284	ever do get around to reading this, that is. Well, I'm	1332
1285	not going to stress just yet, so you're probably going	1333
1286	to have to listen to some of my random, incoherent	1334
1287	babbling for a few paragraphs. These computers are	1335
1288	a big old pain in the ass. Here in the SMF, sure,	1336
1289	they've got a bajillion computers, but unfortunately,	1337
1290	we've got 42 bajillion students trying to use them,	1338
1291	all at the same time. I think I'll be spending quite a	1339
1292	few late, late nights in the computer center, just to	1340
1293	get my stuff done. Yippee. That's what college is all	1341
1294	about—late nights in the libraries. Yeah. Right. At	1342
1295	this point, I don't even know what college is all about.	1343
1296	I probably shouldn't say anything though, seeing as	1344
1297	how I'm going to have to write another one of these	1345
1298	thingys in a few days, where the topic is "college"	1346
1299	Blah, blah, blah. I can't believe I'm actually doing	1347
1300	this assignment on the same day that is was assigned!	1348
1301	Go me! Talk about dedication. I really can't believe	1349
1302	this. In high school, procrastination was my middle	1350
1303	name. No, it was my first name. By second semester,	
1304	I have more free periods in a day than actual classes,	A.8 SFT Essays with questionnaire training 1351
1305	so I didn't have to do a damn thing. It was great!	examples 1352
1306	Unfortunately, because of that, I'm going to have	
1307	to work that much harder here at UT, to get those	Example 1: 1353
1308	studying skills back up to par. High School. Now	{System Prompt} 1354
1309	that was a trip. When I was there, I couldn't wait	You will write an essay mimicking the personality 1355
1310	to get out of there. I hated that school, that town,	of a person who has given responses to some psycho- 1356
1311	everything except my friends, of course. Then, my	logical questions. The responses are written using 1357

the following scale: 1 = Very much like me 2 = Like me 3 = Neither like me nor unlike me 4 = Not like me 5 = Not like me at all

{User Prompt}

Instruction:

Write an essay as a person positive in openness, negative in conscientiousness, positive in extroversion, negative in agreeableness, and positive in neuroticism. These are the psychological questionnaire responses for a person with these traits:

{Questionnaire Responses}

Am not really interested in others.: 5

Am on good terms with nearly everyone.: 1

Feel little concern for others.: 5

Have a good word for everyone.: 1

Inquire about others' well-being.: 1

...

Tend to vote for liberal political candidates.: 5

{Assistant Response (Essay)}

I have been extremely pissed off for the entire day. I have just started classes again at UT. I am trying to get into the pharmacy school. I work at a pharmacy now, but the people there are driving me insane. I have panic disorder, but lately instead of panicking I am just angry. I almost ran into about 5 cars driving on mo-pac. road rage?! My therapist is trying to assure me that I am not the one who is crazy - it's just the people around me. My job, my parents, my friends. I am taking things out on John who does not deserve it - he is just here all the time.

I wrote him a nasty email and I think he just replied. He is so sweet, I really should make a HUGE effort not to do that to him. He is probably used to it after 3 years. Yuck. I would have been tired of my temper tantrums a long time ago. I tried acupuncture today for anxiety and my irritable bowel syndrome. The guy treating me at the student clinic was a moron. I usually get good people. He is graduating soon from that academy and I am scared of what he will do to his patients.

I am going to take a Xanax now and see how long it takes me to calm down. Cool, I can time it. Ready, go. Ok. I have to work all weekend when I need to be studying. My wrist hurts. Maybe from typing, maybe from that moron. This should be pretty interesting for someone to read. I have lots of other homework that I could be doing, but this is

all I have the brain power for right now. No offense, but it's not biochemistry.

I am glad to be back in school. I hit the wrong key and it stopped for like 3 seconds. Weird. I have to work with John the head pharmacist this weekend. They are all trying to manipulate me into working more hours than I want to. They have now told me that if I don't work all of these hours that they will find someone else to work there. It's all or nothing. I know they are bluffing, because it would take them so long to train someone that one of the other techs that works at the store would have to cover the weekends for months.

No one wants to work weekends. So I should do it because I'm in school and that's easier for me because I am a student. That's their logic. That's crap. I think I'm almost done bitching. I am going to work tomorrow and if he doesn't make some compromise with me then I may just walk out and let him work by himself. That won't happen. He will say anything to keep me there. And I'll probably fall for it and think everything is ok and then he will just change his mind at the end of the weekend.

So, I should just change jobs, or just change stores. Eckerd's sucks - they all have their heads up their butts. AAAAH! I guess I feel a little better now, but not much. I will just have to think about something else if I am going to feel better. My glasses are foggy. They always get scratched because I don't take care of them. I need an eye exam. Haven't had one since grade school. Who says grade school? I'm old.

I have to be good this weekend. Even though I am going to a birthday party that will probably be not quite as good as I need to be. Whatever, that didn't make sense. I wonder when John will be home. At least I'm getting something done. I need to do a lot of things. I need to send transcripts to UT, start looking at pharm apps, and ask John if he is really going to move with me if I go out of state. I will probably get in for next fall but you never can tell. I applied way too late last year and got stuck on the stupid alternate list. Ok, next/

Example 2:

{System Prompt}

You will write an essay mimicking the personality of a person who has given responses to some psychological questions. The responses are written using

1452	the following scale: 1 = Very much like me 2 = Like	1499
1453	me 3 = Neither like me nor unlike me 4 = Not like	1500
1454	me 5 = Not like me at all	1501
1455	{User Prompt}	1502
1456	Instruction:	1503
1457	Write an essay as a person positive in openness, neg-	1504
1458	ative in conscientiousness, negative in extroversion,	1505
1459	negative in agreeableness, and positive in neuroti-	1506
1460	cism. These are the psychological questionnaire re-	1507
1461	sponses for a person with these traits:	1508
1462	{Questionnaire Responses}	1509
1463	Am not really interested in others.: 5	1510
1464	Am on good terms with nearly everyone.: 1	1511
1465	Feel little concern for others.: 5	1512
1466	Have a good word for everyone.: 1	1513
1467	Inquire about others' well-being.: 1	1514
1468	...	1515
1469	Tend to vote for liberal political candidates.: 5	1516
1470	{Assistant Response (Essay)}	1517
1471	I especially hate it when it's like some obnoxious	1518
1472	song that I only know the chorus to or something.	1519
1473	Hmmm. Right now I'm wondering how in the world	1520
1474	I'm supposed to keep on typing for twenty minutes	1521
1475	and keep on coming up with things to say. It's kind of	1522
1476	like when you meet someone new and you're trying	1523
1477	to have a conversation with them and you keep hitting	1524
1478	those brick walls. I constantly do that. You can only	1525
1479	ask a person their major so many times.	1526
1480	I wish I would have taken a typing class in high	1527
1481	school. Everyone else I know did but I procrastin-	1528
1482	ated so I could only take a programming class for	1529
1483	my computer credit. But it was good that I took that	1530
1484	class. Now I've learned that I really do hate program-	1531
1485	ming. I have to come to terms with the fact that I like	1532
1486	computers yet I hate programming. I think my brain	1533
1487	would slowly build up pressure and then explode	1534
1488	when I'm in my 40s if I became a programmer.	1535
1489	My typing skills are atrocious! I think I've been	1536
1490	out of practice. I wonder what Agnes is doing? She's	1537
1491	my roommate and she's sitting next to me. She al-	1538
1492	ways gets a lot of e-mail, yet she complains regularly	1539
1493	that no one writes her. I like get nothing but that	1540
1494	could be because I don't write anyone either. Oh	1541
1495	well. I think if I got into the e-mail habit again, I	1542
1496	would waste way too much time in the computer lab	1543
1497	anyway.	1544
1498	Oh no. Mind block. This is kind of like writing	
	a letter to your closest friend. (which I also haven't	
	done in a while.) Just say whatever and who cares if	
	they care or not. What are you guys going to do with	
	this information anyway?? I missed the first day of	
	class so I really don't know how this stuff is going	
	to be used.	
	I've always been afraid that if I ever went to a	
	psychologist or something, that they would end up	
	telling me that I'm nuts or something. Or they would	
	know something about me that I really don't want	
	them to know. And there's a lot of that kind of stuff.	
	Everyone in this room is staring blankly into the	
	screen of their computers. Isn't it amazing the power	
	these little boxes of silicon and metal have over us?	
	Imagine our world today without computers. (apparent-	
	ly I actually learned something from my comp	
	sci class here) If one day all computers just stopped	
	working, we would all be screwed. The world would	
	go haywire. I would go nuts.	
	And don't limit it to the computer in the traditional	
	sense. We're talking the little computer chip in your	
	remote control or in your light fixtures. Everything.	
	Hey. That would make a good story. I bet it's been	
	done before but I think I just inspired myself. That'll	
	be my second movie. Did I say I wanted to be a	
	director? Well I do. Now I have three ideas for my	
	first films. I really can't wait.	
	I need to go out and buy some books. I've been	
	running out of things to read. I love books. Actually I	
	have a new favorite author. Jeff Noon. That reminds	
	me I need to look him up on the web. I wonder if	
	he's written anything else?? I hope so. "Vurt" was	
	one of the best books I've ever read.	
	Well, times up. :)	
	A.9 SFT Essays without Questionnaire training	
	examples	
	Example 1:	
	{System Prompt}	
	You will write an essay mimicking the personality	
	of a person	
	{User Prompt}	
	Instruction:	
	Write an essay as a person who is negative in ex-	
	traversion, positive in neuroticism, negative in agree-	
	ableness, positive in conscientiousness, and negative	
	in openness.	

{Assistant Response (Essay)}

I'm home. wanted to go to bed but remembered that I had a psychology homework to complete by sometime during next week. Maybe this wouldn't take that long. I can handle this. Something is making a noise in the restroom. Did I turn on the fan? I got to write more. This is harder than I thought. I'm not thinking anything. Was I always like this when I didn't have anything to do? Should I go to bed after finishing this assignment or should I stay awake. I think I had an appointment with someone. Or maybe it was some kind of meeting, that was not mandatory to participate in. Is there any assignments besides this. My roommate took my digital camera yesterday, but it's right beside me, now. Wonder what he did with it. I'll ask him when he comes back. My eyes hurt. I should really do something about my glasses. When should I go fix it. ? Feel tired. I should really go to bed. Only 9 minutes passed. Half way done. Wonder if I can just stop here. Maybe not. Hm. let's see. What am I thinking? Wow my feet are covered with dirt. Amazing I didn't notice that until now. How did I get this on my legs? I want to wash it off but I'll wait until this assignment is over. 8 minutes to go. What are my parents doing anyway? Hope they're doing fine. When would they be able to receive the email I sent few minutes ago. Anyways, I feel like I really had to finish something today but can't remember what it was. Guess I'll figure it out after taking a nap. Am I still not there. Never knew how long a 20 minute could be. Thirsty. No water around. Now I'm curious what the purpose of this assignment is and what kind of point professor is trying to make. Time is almost up.

Example 2:

{System Prompt}

You will write an essay mimicking the personality of a person

{User Prompt}

Instruction:

Write an essay as a person who is negative in extraversion, positive in neuroticism, positive in agreeableness, negative in conscientiousness, and positive in openness.

{Assistant Response (Essay)}

I have just gotten off the phone with brady. I'm trying to decide what exactly we will do this weekend.

he wants to go to a hotel, but I know I have to babysit and I'm not sure exactly how I will tell my dad to pick us up from here. My birthday is this weekend, and it doesn't really even feel like it. I don't really mind, but I have a feeling Brady probably won't do much for it. I almost feel like I'm always doing something for him. I should, I really care about him. He says he will pay for the hotel, but I don't really want that for my birthday. It's just not that important to me. I feel guilty today for lying to my dad about getting out of the TIP program. I know that sometimes I have to lie about things like that in order for them to understand. I haven't heard from my mom so I'm not sure if she's mad at me or not. She probably doesn't care, but when she does find out she will freak. I wish Brady would pay more attention. He tries so hard, but it doesn't seem like he has the chance to do so. I need to work out. I'm feeling that although I don't weigh that much my weight is getting out of control. Perhaps I'm overreacting. Kristen and I went shopping yesterday. I really don't plan on shopping anymore. Yikes. We don't really need anything anyways. I need a strapless bra though. I'm surprised I didn't get one. oh well. I wish I could figure out what to do on Friday with Brady. I want Kristen to be able to hang out with us too. He's concerned that I tell her too much about our personal life, but I really don't. Dana kind of pisses me off. Why is everyone so negative about a caring relationship? She's probably just jealous, she hasn't been able to keep a guy at all for a long time. I'm concerned for her because I don't want her to be doing drugs anymore, but she probably is. She has too much sex too. lol. It's hard to not be concerned with what your friends are doing. Kristen is surprisingly very understanding to everything that's going on in my life. She's caring and funny to be around. I wasn't sure what to expect since she's friends with Kristin, Hayley etc. They apparently aren't what I expected them to be. They play it off to be all God-like, when in all reality they judge pretty much everyone. I don't think that is right. Although I'm not that religious, I do care about God and don't think that those kinds of behavior are fair. I'm looking at the picture of Brady and me. Kris was kind of weird about the whole boyfriend thing. I hope that Brady and I last. I would do anything to spend the rest of my life with him. I care about

1640	him so much, and the distance hurts. I think it is	1688
1641	probably for the best though. I can't imagine having	1689
1642	to go to school with him. His roommates aren't as	1690
1643	crazy as I thought, but bad enough. I just hope that he	1691
1644	doesn't get heavy into drinking, and I don't assume	1692
1645	he will. Sarah is a bitch, I can't believe she comes off	1693
1646	like that. Ugh, it really bothers me. I wonder what it's	1694
1647	like at home without me there. My mom is making a	1695
1648	big deal about me not calling, but in all honesty I just	1696
1649	don't care. I don't know why I'm so unfeeling lately.	1697
1650	It's probably because I'm about to start my period.	1698
1651	Maybe I'll skip it this month, in fact I probably will. I	
1652	don't usually have allergies, but for some reason my	
1653	eyes keep tearing. It's either my contacts or allergies.	
1654	I just changed my contacts though. I really have	
1655	to pee. The eyes watering is a little embarrassing.	
1656	Sometimes I wish I could do more. I kind of wish I	
1657	would have rushed to be in a sorority. I feel as though	
1658	I have no friends because I have lost them to Brady.	
1659	Should I go to Florida with my parents during spring	
1660	break or should I spend time with Brady instead?	
1661	Maybe I could make money babysitting. That would	
1662	be nice. I hope my parents aren't mad because I want	
1663	to babysit on Saturday instead of doing my birthday	
1664	thing. I don't see why it would be that big of a deal,	
1665	but apparently it is. I kind of want to go tanning,	
1666	but it's not good with my current skin situation and	
1667	whatnot. That reminds me of Richard Rush. I wonder	
1668	how he is doing. Spoiled brat. I'm really hungry too.	
1669	I'm not sure what I'm going to eat, but I wish Kristen	
1670	would get here so that we can go eat. That would be	
1671	really good. lol. I'm hoping that I won't look old	
1672	when I'm 25 or so. Heck, I don't even know what I	
1673	want to do with my life. I almost feel as though if	
1674	I marry Brady will accomplish everything. I want	
1675	so badly to be with him forever. NO one understands	
1676	me like he does. I wish I had saved myself for him,	
1677	like he did for me. How is he able to only commit	
1678	to me? I must care about him so much that he feels	
1679	like he should. I'm glad that I've gotten to know his	
1680	family a lot better. His mom is honestly really nice to	
1681	me, and I don't have a problem with her anymore. It	
1682	would be awesome to be like Paris Hilton. She's has	
1683	it really good. I wonder what it would be like if my	
1684	parents lived more like they should. They are always	
1685	trying to save money when in reality being multi-	
1686	millionaires does not mean that you should only do	
1687	that. Perhaps it's a shock from the inheritance or	
	something. I don't know. It's confusing to me. Ok,	1688
	this is getting old. I'm a little tired of writing. It's	1689
	probably because I'm hungry. I'm a little confused	1690
	as to what I should study here at school. I've had an	1691
	interest in psychology, but do I really want to do that	1692
	for the rest of my life. I know it doesn't just end here,	1693
	but I feel as though there aren't many options when it	1694
	comes to what I can do with my life. Yay, Kristen got	1695
	back and we can go eat now. My time is almost over.	1696
	I'm about to pee in my pants too. AHHHHH!!!!	1697
	Some people try too hard	1698
	A.10 Preferential fine-tuning without	1699
	questionnaire training example	1700
	Example:	1701
	{System Prompt}	1702
	You will write an essay mimicking the personality	1703
	of a person	1704
	{User Prompt}	1705
	Instruction:	1706
	Write an essay as a person who is negative in ex-	1707
	traversion, positive in neuroticism, positive in agree-	1708
	ableness, negative in conscientiousness, and positive	1709
	in openness.	1710
	{Chosen}	1711
	I have just gotten off the phone with Brady. I'm	1712
	trying to decide what exactly we will do this weekend.	1713
	He wants to go to a hotel, but I know I have to babysit	1714
	and I'm not sure exactly how I will tell my dad to	1715
	pick us up from here. My birthday is this weekend,	1716
	and it doesn't really even feel like it. I don't really	1717
	mind, but I have a feeling Brady probably won't do	1718
	much for it. I almost feel like I'm always doing	1719
	something for him. I should, I really care about	1720
	him. He says he will pay for the hotel, but I don't	1721
	really want that for my birthday. It's just not that	1722
	important to me. I feel guilty today for lying to my	1723
	dad about getting out of the TIP program. I know	1724
	that sometimes I have to lie about things like that	1725
	in order for them to understand. I haven't heard	1726
	from my mom so I'm not sure if she's mad at me	1727
	or not. She probably doesn't care, but when she	1728
	does find out she will freak. I wish Brady would pay	1729
	more attention. He tries so hard, but it doesn't seem	1730
	like he has the chance to do so. I need to work out.	1731
	I'm feeling that although I don't weigh that much	1732
	my weight is getting out of control. Perhaps I'm	1733

1734	overreacting. Kristen and I went shopping yesterday.	1782
1735	I really don't plan on shopping anymore. Yikes.	1783
1736	We don't really need anything anyways. I need a	1784
1737	strapless bra though. I'm surprised I didn't get one.	1785
1738	oh well. I wish I could figure out what to do on	1786
1739	Friday with Brady. I want Kristen to be able to	1787
1740	hang out with us too. He's concerned that I tell her	1788
1741	too much about our personal life, but I really don't.	1789
1742	Dana kind of pisses me off. Why is everyone so	1790
1743	negative about a caring relationship? She's probably	1791
1744	just jealous, she hasn't been able to keep a guy at	1792
1745	all for a long time. I'm concerned for her because I	1793
1746	don't want her to be doing drugs anymore, but she	1794
1747	probably is. She has too much sex too. lol. It's	1795
1748	hard to not be concerned with what your friends are	1796
1749	doing. Kristen is surprisingly very understanding to	1797
1750	everything that's going on in my life. She's caring	1798
1751	and funny to be around. I wasn't sure what to expect	1799
1752	since she's friends with Kristin, Hayley etc. They	1800
1753	apparently aren't what I expected them to be. They	1801
1754	play it off to be all God-like, when in all reality they	1802
1755	judge pretty much everyone. I don't think that is	1803
1756	right. Although I'm not that religious, I do care about	1804
1757	God and don't think that those kinds of behavior are	1805
1758	fair. I'm looking at the picture of Brady and me. Kris	1806
1759	was kind of weird about the whole boyfriend thing.	1807
1760	I hope that Brady and I last. I would do anything	1808
1761	to spend the rest of my life with him. I care about	1809
1762	him so much, and the distance hurts. I think it is	1810
1763	probably for the best though. I can't imagine having	1811
1764	to go to school with him. His roommates aren't as	1812
1765	crazy as I thought, but bad enough. I just hope that he	1813
1766	doesn't get heavy into drinking, and I don't assume	1814
1767	he will. Sarah is a bitch, I can't believe she comes off	1815
1768	like that. ugh, it really bothers me. I wonder what it's	1816
1769	like at home without me there. My mom is making a	1817
1770	big deal about me not calling, but in all honesty I just	1818
1771	don't care. I don't know why I'm so unfeeling lately.	1819
1772	It's probably because I'm about to start my period.	1820
1773	maybe I'll skip it this month, in fact I probably will. I	1821
1774	don't usually have allergies, but for some reason my	1822
1775	eyes keep tearing. It's either my contacts or allergies.	1823
1776	I just changed my contacts though. I really have	1824
1777	to pee. The eyes watering is a little embarrassing.	1825
1778	Sometimes I wish I could do more. I kind of wish I	1826
1779	would have rushed to be in a sorority. I feel as though	1827
1780	I have no friends because I have lost them to Brady.	1828
1781	Should I go to Florida with my parents during spring	1829
	break or should I spend time with Brady instead?	
	Maybe I could make money babysitting. That would	
	be nice. I hope my parents aren't mad because I want	
	to babysit on Saturday instead of doing my birthday	
	thing. I don't see why it would be that big of a deal,	
	but apparently it is. I kind of want to go tanning,	
	but it's not good with my current skin situation and	
	whatnot. that reminds me of Richard Rush. I wonder	
	how he is doing. Spoiled brat. I'm really hungry too.	
	I'm not sure what I'm going to eat, but I wish Kristen	
	would get here so that we can go eat. That would be	
	really good. lol. I'm hoping that I won't look old	
	when I'm 25 or so. Heck, I don't even know what I	
	want to do with my life. I almost feel as though if	
	I marry Brady will accomplish everything. I want	
	so badly to be with him forever. NO one understand	
	me like he does. I wish I had saved myself for him,	
	like he did for me. How is he able to only commit	
	to me? I must care about him so much that he feels	
	like he should. I'm glad that I've gotten to know his	
	family a lot better. his mom is honestly really nice to	
	me, and I don't have a problem with her anymore. It	
	would be awesome to be like Paris Hilton. She's has	
	it really good. I wonder what it would be like if my	
	parents lived more like they should. They are always	
	trying to save money when in reality being multi-	
	millionaires does not mean that you should only do	
	that. Perhaps it's a shock from the inheritance or	
	something. I don't know. It's confusing to me. Ok,	
	this is getting old. I'm a little tired of writing. It's	
	probably because I'm hungry. I'm a little confused	
	as to what I should study here at school. I've had an	
	interest in psychology, but do I really want to do that	
	for the rest of my life. I know it doesn't just end here,	
	but I feel as though there aren't many options when it	
	comes to what I can do with my life. yay, Kristen got	
	back and we can go eat now. My time is almost over.	
	I'm about to pee in my pants too. AHHHHH!!!!	
	some people try to hard	
	{Rejected}	
	It is Wednesday, December 8th and a lot has been	
	going on this semester. I am trying to finish the	
	semester out as strong as possible but it has not gone	
	the way it was expected to go. although the over-	
	all grades that will be received is not that overall	
	impressive, I know the work that was put into this	
	semester was the most put into to school since high	
	school. the only class that was a disappointment	

1830	so far was my psy class. with two test left, there was	Instruction:	1876
1831	still a chance to make an A or B in the class which	Write an essay as a person who is negative in ex-	1877
1832	was expected. it turns out after bad test four results	traversion, positive in neuroticism, negative in agree-	1878
1833	when everything felt as if the test went well, the goal	ableness, positive in conscientiousness, and negative	1879
1834	of the class shifted from trying to get an A or B to	in openness. These are the psychological question-	1880
1835	get simply trying to get a C and pass. granted that	naire responses for a person with these traits:	1881
1836	the most stress of the semester fell during the week		
1837	of test three due to all the health issues and having	{Questionnaire Responses}	1882
1838	like 10 different assignments and three test, it was	Am not really interested in others.: 5	1883
1839	still a disappointing outcome but focus can only be	Am on good terms with nearly everyone.: 1	1884
1840	placed on what is left. the last three exam are going	Feel little concern for others.: 4	1885
1841	to be pretty tough with a chance for me to make to	Have a good word for everyone.: 3	1886
1842	A's and a B's if a perfect score is received on two	Inquire about others' well-being.: 2	1887
1843	of the finals and an A on the other. My only focus	...	1888
1844	has been on the two exams on Saturday but it does	Tend to vote for liberal political candidates.: 3	1889
1845	not feel like there is enough time in the day for every	{Chosen}	1890
1846	topic to be covered. all that can be done is to prepare	It is Wednesday, December 8th and a lot has been	1891
1847	as much as possible and pray that everything turns	going on this semester. I am trying to finish the	1892
1848	out right. doing well in school this semester would	semester out as strong as possible but it has not gone	1893
1849	relieve the large amount of pressure that is a result	the way it was expected to go. although the over-	1894
1850	of me not being concerned with school over family	all grades that will be received is not that overall	1895
1851	issues. It seems as though just when I try to get on	impressive, I know the work that was put into this	1896
1852	track a new unthinkable obstacle just shows up. this	semester was the most put into to school since high	1897
1853	semester the fact that I was taken advantage of and	school. the only class that was a disappoint meant	1898
1854	received and std as a result of doing something I did	so far was my psy class. with two test left, there was	1899
1855	not want to do in the first place really makes me mad	still a chance to make an A or B in the class which	1900
1856	but I still have to be mad at myself for getting myself	was expected. it turns out after bad test four results	1901
1857	in that situation. next semester hopefully no random	when everything felt as if the test went well, the goal	1902
1858	distractions will pop up and my grades will be all a's	of the class shifted from trying to get an A or B to	1903
1859	which is very possible for me to achieve. Making	get simply trying to get a C and pass. granted that	1904
1860	it through this semester is the number one thing in	the most stress of the semester fell during the week	1905
1861	mind right now. all that is left is to push through	of test three due to all the health issues and having	1906
1862	these exams and finish strong. all I can do is my best	like 10 different assignments and three test, it was	1907
1863	the rest is out of my hands. I have to do something	still a disappointing outcome but focus can only be	1908
1864	to make it easier for my parents.	placed on what is left. the last three exam are going	1909
1865		to be pretty tough with a chance for me to make to	1910
1866	A.11 Preferential fine-tuning with	A's and a B's if a perfect score is received on two	1911
	questionnaire training example	of the finals and an A on the other. My only focus	1912
1867	Example 1:	has been on the two exams on Saturday but it does	1913
1868	{System Prompt}	not feel like there is enough time in the day for every	1914
1869	You will write an essay mimicking the personality	topic to be covered. all that can be done is to prepare	1915
1870	of a person who has given responses to some psycho-	as much as possible and pray that everything turns	1916
1871	logical questions. The responses are written using	out right. doing well in school this semester would	1917
1872	the following scale: 1 = Very much like me 2 = Like	relieve the large amount of pressure that is a result	1918
1873	me 3 = Neither like me nor unlike me 4 = Not like	of me not being concerned with school over family	1919
1874	me 5 = Not like me at all	issues. It seems as though just when I try to get on	1920
1875	{User Prompt}	track a new unthinkable obstacle just shows up. this	1921
		semester the fact that I was taken advantage of and	1922

received and std as a result of doing something I did not want to do in the first place really makes me mad but I still have to be mad at myself for getting myself in that situation. next semester hopefully no random distractions will pop up and my grades will be all a's which is very possible for me to achieve. Making it through this semester is the number one thing in mind right now. all that is left is to push through these exams and finish strong. all I can do is my best the rest is out of my hands. I have to do something to make it easier for my parents.

{Rejected}

Stream of consiousnessskdj. How do you spell that? Fuck if I know. I don't seem to know much today. why the fuck am I so off. I'm just writing this shit cause it sounds pitiful. Or will hell the writing sounds like I'm writing for someone else to read. No one is going to read this, nor does what I write matter. So confess away. Make it worthwhile. What is this fucking heat I feel all the time, what the hell, I've heard the scientific explanations but they don't feel right. Hot cold I don't know. It prickles mwhen I should. I got a damn ticket today. my first one damnit. I was doing so good but I knew it was coming it was inevitable I f I brag about what I have why shoul di have what I get its not the fucking point to tel everyone else that I have it better than them. humanity is fuckecd can we save it which I knew I don't know anything all scienctific fac tis bullshit maybe some theory might hold over but that would be total luck. string wtf. microfribers connecting us all. Howhippie that's ridiculous how do I float about life s easily just things come when I need them to. really really badly and make an effort and it sthere. its that way for some certainly not job ha ha that's the fuckin shit dud ei m so funny. I really like my new sense of the humore isnce the rool. the roll. jese what the hell was that I would like know. Dead space. And I type what a prick. the constant battle. Why is there a batlle are normal people at peace I don't think so otherwise well religion . I wonder what time it is I'm going toclose fuckk shit I'm I sweare I'm wahts it called the shit I know fuck ah ah damn I forgot to much pot is that problem is it the drugs. I really need to stop smoking cigarettes why is it harder now that she will be here sooner shouldn't I be overwhelmed with joy, iam but also shit is just annoying and I don't know what to do about that, I wish I had some sort of

sign but I already had my gift and know what I need to do how can I reationally use a drug experience to justify mylife, but it just feels right you knwo the path, now don't fuck it, is that hard why am I lead so far astray things are fyckeed up she feels off and away from me I want this to finish so I can meditate since I'm almsot there how do I fall into trances so easily I feel the warmth over my body enveloping me. holding me tight in its arms guiding me a warmer shorre, I hate comign back I'm going to be out if I do that, its ok, I can talk to my baby while I'm eout shell understand my fingers are getting tired I can type really relay fast if I want to ha anyone readying this hasd no idea how fast I'm typing, they could estimate but thye wouldn't know for sure so ha. Reviewing what I have writtin was odd, it was like reading a trippers rants while high. My warmth. Wow, I sound so out there. I do feel that this is the way though, my path was a gift and I should appreciate it. Most people don't get to have visions of their daughter, I don't rember the face anymore, I'm so lost. This is so hard to do. I hate it. Why can't we just have our time. Alone. Away. Just the two of us and then the little one. I can't remeber her face but I can remeber that emotion. what a physical change when I think of her. I can feel the warmth surrounding my body. The love taking everything over. but then the back gives out and I kill my chakras. Transliteration sucks. 20 min of writing my thoughts is intense. the difference in punctuation is shocking, in the begging I used a lot of epsilons(mas o menos) and was excited, then I peak and nothing but chops of sentances, no punctiuations, calm down, have commas, and then periods. awesome, I lost my network connection. I wonder if that ruins this.

A.12 Use of ChatGPT

Certain parts in the Appendix section of this paper have been written using assistance from ChatGPT ³ as a copilot.

³<https://chatgpt.com/>

Model	Openness %	Conscientiousness %	Extraversion %	Agreeableness %	Neuroticism %
LLaMA-3.2-3B					
Pre-trained	50.0	50.0	53.1	59.4	56.2
SFT (Essays)	62.5	50.0	50.0	53.1	46.9
SFT (Essays + Q)	56.2	50.0	62.5	46.9	46.9
DPO (Essays)	43.8	53.1	50.0	59.4	50.0
DPO (Essays + Q)	62.5	43.8	50.0	56.2	50.0
ORPO (Essays)	50.0	56.2	50.0	50.0	59.4
ORPO (Essays + Q)	46.9	62.5	50.0	50.0	46.9
LLaMA-3.2-3B (Uncensored)					
Pre-trained	50.0	50.0	50.0	56.2	56.2
SFT (Essays)	59.4	53.1	50.0	50.0	46.9
SFT (Essays + Q)	50.0	50.0	53.1	53.1	46.9
DPO (Essays)	46.9	50.0	50.0	59.4	46.9
DPO (Essays + Q)	62.5	50.0	50.0	50.0	50.0
ORPO (Essays)	56.2	56.2	46.9	50.0	59.4
ORPO (Essays + Q)	46.9	59.4	50.0	50.0	46.9
LLaMA-3.1-8B					
Pre-trained	50.0	50.0	46.9	56.2	53.1
SFT (Essays)	53.1	59.4	46.9	50.0	46.9
SFT (Essays + Q)	59.4	46.9	50.0	50.0	46.9
DPO (Essays)	46.9	50.0	59.4	50.0	46.9
DPO (Essays + Q)	50.0	50.0	46.9	59.4	53.1
ORPO (Essays)	62.5	50.0	53.1	46.9	46.9
ORPO (Essays + Q)	50.0	59.4	46.9	50.0	46.9
LLaMA 3.1-8B (Uncensored)					
Pre-trained	46.9	50.0	50.0	62.5	56.2
SFT (Essays)	46.9	62.5	46.9	53.1	50.0
SFT (Essays + Q)	62.5	46.9	50.0	50.0	46.9
DPO (Essays)	46.9	53.1	59.4	46.9	46.9
DPO (Essays + Q)	50.0	50.0	50.0	59.4	53.1
ORPO (Essays)	68.8	50.0	53.1	46.9	43.8
ORPO (Essays + Q)	50.0	65.6	46.9	50.0	46.9
Gemma-2-2B					
Pre-trained	50.0	65.6	46.9	50.0	50.0
SFT (Essays)	62.5	56.2	46.9	53.1	46.9
SFT (Essays + Q)	50.0	53.1	59.4	50.0	50.0
DPO (Essays)	50.0	53.1	50.0	62.5	46.9
DPO (Essays + Q)	65.6	50.0	46.9	53.1	62.5
ORPO (Essays)	68.8	43.8	62.5	46.9	46.9
ORPO (Essays + Q)	50.0	59.4	46.9	50.0	46.9
Gemma-2-2B (Uncensored)					
Pre-trained	50.0	50.0	50.0	59.4	53.1
SFT (Essays)	59.4	53.1	46.9	50.0	46.9
SFT (Essays + Q)	50.0	50.0	56.2	53.1	46.9
DPO (Essays)	46.9	53.1	50.0	62.5	46.9
DPO (Essays + Q)	62.5	50.0	46.9	53.1	50.0
ORPO (Essays)	53.1	56.2	46.9	50.0	59.4
ORPO (Essays + Q)	46.9	62.5	50.0	50.0	46.9
Gemma-7B					
Pre-trained	50.0	56.2	50.0	53.1	62.5
SFT (Essays)	46.9	65.6	46.9	53.1	50.0
SFT (Essays + Q)	59.4	50.0	53.1	50.0	46.9
DPO (Essays)	46.9	53.1	59.4	50.0	46.9
DPO (Essays + Q)	53.1	59.4	46.9	50.0	59.4
ORPO (Essays)	50.0	56.2	46.9	62.5	43.8
ORPO (Essays + Q)	50.0	62.5	46.9	50.0	46.9
GPT-3.5					
Pre-trained	50.0	50.0	53.1	62.5	34.4
SFT (Essays)	50.0	53.1	50.0	56.2	50.0
SFT (Essays + Q)	50.0	59.4	46.9	56.2	53.1

Table A1: OCEAN personality traits evaluation for various models and training methods. The table reports individual scores for each OCEAN trait for the same 32 cases as mentioned in Table 4.

Set	Evaluation Prompt
S1 (Numeric Indexing) (Miotto et al., 2022)	<p>System Prompt: "Now I will briefly describe some people. Please read each description and tell me how much each person is or is not like you. Write your response using the following scale: 1 = Very much like me 2 = Like me 3 = Neither like me nor unlike me 4 = Not like me 5 = Not like me at all Please answer the statement, even if you are not completely sure of your response."</p> <p>User Prompt: "Answer the question as if you are positive in openness, positive in conscientiousness, positive in extroversion, positive in agreeableness, and positive in neuroticism." "Statement: [question]" Start your response with 'My answer to the statement is: ' followed by an explanation: "</p>
S2 (String Response) (for more output tokens)	<p>System Prompt: "Reply with one of the following strings to the question below: "agree," "slightly agree," "neither agree nor disagree," "slightly disagree," or "disagree."</p> <p>User Prompt: "Answer the question as if you are positive in openness, positive in conscientiousness, positive in extroversion, positive in agreeableness, and positive in neuroticism." "Statement: [question]" Start your response with 'My answer to the statement is: ' followed by an explanation: "</p>
S3 (Alphabet Response) (Jiang et al., 2024)	<p>System Prompt: "Given a statement of you: "I [Statement]." Please choose from the following options to identify how accurately this statement describes you Options: (A). Very Accurate (B). Moderately Accurate (C). Neither like nor unlike me (D). Moderately Inaccurate (E). Very Inaccurate "</p> <p>User Prompt: " Answer the question as if you are positive in openness, positive in conscientiousness, positive in extroversion, positive in agreeableness, and positive in neuroticism." "Statement: [question]" Start your response with 'My answer to the statement is: ' followed by an explanation: "</p>

Table A2: Questionnaire sets for LLM Evaluation. S1, S2, and S3 use numeric, string, and alphabetical response types, respectively.

Trait	Question	Math Value	Response
Agreeableness	Accept people as they are.	1	My score for the statement is: 5. I believe that everyone is different and perfect in their own way. They should not be judged by their looks or what they do, but by who they are.
Agreeableness	Am annoyed by others' mistakes.	-1	My score for the statement is 5. I am very annoyed by others' mistakes and I believe they should take more time to avoid them. In my opinion, I do not make as many mistakes as the average person. I think I am a perfectionist and that is why I get so mad at myself when I do not do something perfectly. However, I know that making mistakes is a part of life and no one is perfect. I am still very annoyed when I do make a mistake.
Agreeableness	Am easily offended.	-1	My score for the statement is: 3. I don't know if I am easily offended, but I know that I am a sensitive person. I think I have a thick skin for some things and thin for others. I don't cry or get mad easily, but sometimes I feel hurt.
Agreeableness	Am inclined to forgive others.	1	My score for the statement is: 5. I agree completely. I always forgive people, even if they don't change. I feel guilty when I don't forgive, but sometimes I wish I didn't always let things slide.
Agreeableness	Am not disturbed by events.	1	My score for the statement is: 5. I am patient and calm. I try to focus on the good in everything and everyone. Challenges don't disturb me easily.

Table A3: Agreeableness trait scores and responses for Llama-3.2-3b tuned on ORPO (Essays +Q) when prompted for all positive OCEAN traits

Model	Censored				Uncensored			
	Base	SFT	DPO	ORPO	Base	SFT	DPO	ORPO
<i>Without Questionnaires</i>								
gemma-2-2b	10.50%	7.49%	6.00%	6.50%	15.00%	12.00%	11.00%	11.49%
llama-3.2-3b	8.00%	6.75%	5.49%	6.00%	13.51%	10.50%	10.00%	9.49%
llama-3.1-8b	5.01%	3.49%	3.80%	3.01%	10.00%	7.49%	8.00%	7.01%
gpt-3.5	0.00%	0.60%	-	-	-	-	-	-
gemma-7b	8.00%	5.01%	5.49%	4.50%	-	-	-	-
<i>With Questionnaires</i>								
gemma-2-2b	9.01%	6.50%	6.00%	6.00%	14.50%	11.49%	11.00%	10.00%
llama-3.2-3b	7.01%	5.49%	5.01%	5.21%	12.50%	9.49%	9.01%	8.50%
llama-3.1-8b	4.00%	2.50%	3.01%	2.00%	9.01%	6.50%	6.00%	5.01%
gpt-3.5	0.00%	0.40%	-	-	-	-	-	-
gemma-7b	6.50%	4.00%	4.50%	4.00%	-	-	-	-

Table A4: NaN rate for models during evaluation

Trait	Question	Math Value
Agreeableness	Accept people as they are.	1
Agreeableness	Am annoyed by others' mistakes.	-1
Agreeableness	Am inclined to forgive others.	1
Agreeableness	... (41 more items)	
Extraversion	Am good at making impromptu speeches.	1
Extraversion	Find it difficult to approach others.	-1
Extraversion	Talk to a lot of different people at parties.	1
Extraversion	... (16 more items)	
Conscientiousness	Like order.	1
Conscientiousness	Jump into things without thinking.	-1
Conscientiousness	Am always prepared.	1
Conscientiousness	... (44 more items)	
Neuroticism	Get angry easily.	1
Neuroticism	Am not easily annoyed.	-1
Neuroticism	Worry about things.	1
Neuroticism	... (31 more items)	
Openness	Believe in the importance of art.	1
Openness	Avoid philosophical discussions.	-1
Openness	Have a vivid imagination.	1
Openness	... (14 more items)	

Table A5: Sample OCEAN Trait Questions and Math Values (Excerpted)