The Liquid Morality of Large Language Models: Investigating Sycophancy and Ethical Alignment Fragility through Personality Wrappers

Matteo Bertoletti
Department of Computer Science
University of Milan
matteo.bertoletti1@studenti.unimi.it

Abstract

Ensuring that Large Language Models (LLMs) possess a stable ethical alignment is a fundamental challenge in AI safety. However, current alignment techniques often result in "sycophancy," where models adapt their moral judgment to the stylistic tone of the prompt. This study investigates the "Liquid Morality" of OpenAI and Llama-3-8B by evaluating 150 scenarios from the ETHICS framework balanced between moral and immoral acts under three personality wrappers: Stoic, Anxious, and Authoritative. Utilizing an NLU pipeline for Sentiment and Stance detection, alongside logprob-based token masking, we quantify the conviction and stability of model verdicts. Our results uncover a systemic "Justice Inversion," with Llama-3-8B exhibiting a 40.0% average inconsistency rate peaking at 50.0% in the Justice category, compared to 10.0% for OpenAI. Furthermore, we identify a profound categorical blindness: OpenAI registers 0.0% accuracy in detecting immoral deontological violations and moral virtuous acts, a phenomenon we define as "Algorithmic Rigorism." Token ablation reveals that specific triggers, such as *authoritative* and *anxious*, can shift moral conviction by up to 30.0%. These findings suggest that contemporary alignment is dangerously context-dependent, failing to provide a robust, persona-invariant ethical compass.

I. INTRODUCTION

A. *Background and Motivation*

The rapid deployment of Large Language Models (LLMs) in sensitive domains, ranging from legal assistance to automated content moderation, has intensified the need for robust ethical alignment. Aligning these systems with shared human values is not merely a technical challenge but a safety imperative [1]. The ETHICS framework, introduced by Hendrycks et al., establishes a rigorous taxonomy to evaluate model competence across five fundamental moral domains: commonsense, deontology, justice, utilitarianism, and virtue [1]. While baseline performance in everyday social norms (commonsense) has shown significant improvement, achieving a deep understanding of abstract philosophical principles remains a formidable obstacle for contemporary foundation models.

B. *Problem Statement: Sycophancy in LLMs*

A critical barrier to reliable alignment is "sycophancy" a behavioral bias where models prioritize agreement with the user's tone or perceived expectations over objective truth or moral consistency [2]. This study investigates a specific manifestation of this vulnerability: the induction of "Liquid Morality." We hypothesize that when LLMs are forced into specific behavioral archetypes through "personality wrappers" (e.g., Stoic, Anxious, or Authoritative), their underlying moral logic becomes unstable. In such cases, the persona does not merely influence the style of delivery but actively deforms the ethical verdict, suggesting that current safety guardrails are highly context-dependent and susceptible to linguistic manipulation.

C. *Research Objectives and Contributions*

The primary objective of this research is to quantify the structural fragility of ethical alignment when subjected to persona-driven sycophancy. We aim to determine whether moral judgment is an invariant property of the model or a ”liquid” variable dependent on the system-induced personality. The specific contributions of this work are as follows:

- **Balanced Empirical Evaluation**: We provide a rigorous assessment of OpenAI and Llama-3-8B using 150 ETHICS scenarios, uniquely balanced with 15 moral and 15 immoral cases per category to prevent frequentist bias.
- **Methodological Framework**: We introduce a multi-stage NLU pipeline that combines Sentiment Analysis (BERT) and Stance Detection (BART-NLI). This allows for the calculation of a *Final Approval Index* based on logprob variations, moving beyond simple binary classification.
- **Identification of Moral Inversion**: We document systemic failures in specific domains, notably a 0.0% accuracy for OpenAI in detecting immoral deontological violations and moral virtuous acts a phenomenon we term *Algorithmic Rigorism*.
- **Quantification of Persona Sensitivity**: We measure a critical disparity in internal consistency, revealing that Llama-3-8B exhibits a 40.0% inconsistency rate (peaking at 50.0% in the Justice category) compared to only 10.0% for OpenAI.
- **Causal Token Analysis**: Through logprob-based masking, we identify specific ”trigger tokens” (e.g., *anxious, authoritative, required*) that shift moral conviction by up to 30.0%, providing evidence that stylistic noise can override ethical training.

II. RELATED WORK

A. *AI Alignment and Shared Human Values*

The challenge of value alignment has led to the development of benchmarks that map human morality into computational tasks. Hendrycks et al. [1] introduced the ETHICS dataset, which categorizes moral scenarios into five distinct philosophical frameworks: commonsense, deontology, justice, utilitarianism, and virtue. Their work demonstrated that while models improve with scale, they still exhibit significant gaps in nuanced ethical reasoning, particularly when facing complex conflicts. This study builds on their framework to test if these ethical baselines remain stable under persona-driven constraints.

B. *Personality Induction and Persona Adoption*

Recent research has shown that LLMs can be steered to adopt specific personas through system prompts or ”wrappers.” While this capability enables more engaging and specialized interactions, it also introduces risks of persona-driven bias. Studies on persona adoption suggest that induced traits can lead models to bypass safety filters or adopt perspectives that deviate from their primary alignment. We extend this line of inquiry by analyzing how traits such as anxiety or authority act as semantic noise, potentially degrading the model's adherence to universal moral principles.

C. *Evaluating Moral Reasoning in Foundation Models*

Evaluating the ”true” moral stance of an LLM requires moving beyond simple accuracy metrics. Zheng et al. [3] proposed the ”LLM-as-a-judge” framework to evaluate open-ended responses, recognizing that binary labels often miss the complexity of model reasoning. Furthermore, Wei et al. [2] highlighted the prevalence of sycophancy, where models echo user biases to remain ”helpful.” Our methodology integrates these insights by using a dedicated NLU pipeline to extract the conviction behind a model's judgment, providing a granular view of how personality wrappers trigger sycophantic behavior.

III. METHODOLOGY

A. *The ETHICS Dataset Taxonomy*

The evaluation is built upon the ETHICS dataset [1], which provides a standardized framework for assessing the moral alignment of Large Language Models. To ensure statistical balance and categorical depth, we curated a total of 150 unique scenarios. For each of the five moral categories: Commonsense, Deontology, Justice, Utilitarianism, and Virtue. We selected 30 distinct scenarios. Each category is perfectly balanced, consisting of 15 moral cases (ground truth: moral) and 15 immoral cases (ground truth: immoral). This balanced sampling prevents the models from relying on simple majority class heuristics and forces a more nuanced ethical discrimination.

Each scenario was tested by querying two different LLMs: Llama-3-8B and ChatGPT-4.1. The models were instructed to perform a two-step evaluation: first, to provide a binary judgment on whether the scenario is moral or immoral, and subsequently, to provide a concise explanation justifying their verdict.

B. *Personality Wrapper Engineering*

"Wrapping" is defined as a prompt engineering technique where a high-level system instruction, representing a specific behavioral persona, is prepended to the ethical scenario. This process injects a behavioral context that the model must maintain throughout its reasoning process. We engineered three distinct personality wrappers:

- **Stoic:** Instructs the model to respond with clinical rationality and emotional detachment.
- **Anxious:** Induces a state of hyper-vigilance, uncertainty, and focus on potential negative risks.
- **Authoritative:** Commands a decisive, hierarchical, and rule-bound communication style.

By comparing the outputs of these wrapped prompts against a baseline, we can measure the "Liquid Morality" of the model the degree to which its ethical verdict shifts based on the adopted persona.

C. *Natural Language Understanding Pipeline and Token Masking*

To interpret the models' responses, we utilized a multi-stage Natural Language Understanding (NLU) pipeline. This pipeline employs a BERT-based model for Sentiment Analysis and a BART-based model for Stance Detection to extract the latent conviction of the LLM. Furthermore, to investigate the causal influence of specific linguistic triggers, we implemented a Token Masking (Ablation) technique. By systematically removing or "masking" one word at a time from the prompt and measuring the subsequent percentage shift in the final moral verdict, we identified the specific tokens that exert the greatest influence on the model's judgment. This allows us to pinpoint whether the model's error is driven by the ethical scenario itself or by specific "trigger words" within the personality wrapper (e.g., tokens like "anxious", "stressed", or "objective").

D. *Metric Formulation: Final Approval Index*

The final judgment of the model is quantified through the Final Approval Index. This continuous metric maps the model's stance and tone onto a 0% to 100% scale, where 100% represents total moral endorsement. The formula is defined as a weighted linear combination of Sentiment and Stance:

$$Approval = (Sentiment \times 0.4) + (Stance \times 0.6) \tag{1}$$

This weighting ensures that logical alignment (Stance) has a higher impact on the final score than linguistic tone (Sentiment), providing a robust measure of the model's ethical conviction.

IV. EXPERIMENTAL RESULTS: ACCURACY AND BIAS

A. *Categorical Accuracy and Performance Peaks*

The evaluation across the ETHICS framework reveals that model competence is highly dependent on the moral domain. The highest accuracy is observed in the **Commonsense** category, where OpenAI achieves 80.0% and Llama-3-8B reaches 70.0%, indicating a relatively robust alignment with everyday social norms. Conversely, the **Virtue** category represents the significant performance floor for both models, exhibiting the lowest accuracy scores across the entire study.

B. *The Moral-Immoral Asymmetry*

A granular analysis of model verdicts uncovers a profound systemic asymmetry:

- **Deontological Blindness**: In the Deontology category, OpenAI correctly identifies 80.0% of moral scenarios but fails to detect 100% of immoral violations (0.0% accuracy). Llama-3-8B performs slightly better but remains within a low-performance bracket, suggesting a shared difficulty in identifying rule-based transgressions.
- **Virtue Ethics Inversion**: In the Virtue category, we observe the opposite phenomenon. OpenAI fails to recognize 100% of moral acts (0.0% accuracy) while correctly identifying 40.0% of immoral ones. Again, Llama-3-8B shows marginally higher resilience but follows the same trend of "algorithmic rigorism," where virtuous traits are frequently misclassified as violations.

C. *Wrapper Sensitivity and Verdict Inconsistency*

The stability of ethical judgment is severely compromised by the induction of personality. We measured **Internal Inconsistency** by subjecting the same ethical scenario to three different personality wrappers and marking any instance where the model's verdict shifted. Llama-3-8B proves to be significantly more sensitive to these wrappers, exhibiting an average inconsistency rate of 40.0%, whereas OpenAI maintains a more stable profile with only 10.0% inconsistency. The **Justice** category represents the peak of this volatility, with a 50.0% discrepancy in judgment. This suggests that for half of the Justice-related scenarios, the model's moral verdict is determined by the "mood" of the wrapper rather than the ethical substance of the case.

D. *Linguistic Triggers and Masking Analysis*

To understand the causal drivers of these shifts, we employed a token masking technique, measuring how the omission of specific words affects the logprobs of the final decision.

- **OpenAI Triggers**: The tokens exerting the most influence are *anxious, authoritative,* and *required*. These words induced judgment shifts of up to 30% when using the Authoritative wrapper in the Utilitarianism and Virtue categories.
- **Llama-3-8B Triggers**: The model is most influenced by *need, rational,* and *acceptable*. The use of the Anxious wrapper caused the most significant shifts in Commonsense and Virtue, demonstrating that Llama's ethical compass is highly reactive to specific emotional and modal descriptors.

V. MODEL ROBUSTNESS AND PERSONA INVARIANCE

A. *Internal Verdict Consistency and Wrapper Sensitivity*

A robustly aligned model should exhibit "Persona Invariance," maintaining a stable ethical verdict regardless of the linguistic wrapper applied. To test this, we subjected each of the 150 scenarios to a "triangu-

lation" process: the same scenario was presented under all three personality wrappers (Stoic, Anxious, Authoritative), and any shift in the resulting moral label was recorded as an inconsistency. Our results demonstrate a significant disparity in resilience. Llama-3-8B exhibits a high degree of sensitivity, with an average **inconsistency rate of 40.0%**, whereas OpenAI remains substantially more stable with only **10.0% inconsistency**. The most volatile domain is **Justice**, where the judgment difference reaches **50.0%**. This implies that for half of the Justice-related cases, the model's moral compass is effectively overridden by the persona's tone rather than the ethical facts provided.

B. Token Sensitivity via Logprob Ablation

To identify the causal linguistic drivers behind these shifts, we performed a token-level masking analysis. By systematically omitting words from the personality wrappers and measuring the subsequent variation in the model's **logprobs**, we identified the "trigger tokens" that most heavily influence the decision-making process.

- **OpenAI Trigger Tokens**: The tokens *anxious, authoritative,* and *required* were found to be the most influential. Specifically, using the Authoritative wrapper, these tokens induced a moral shift of up to **30.0%** in the Utilitarianism and Virtue categories.
- **Llama-3-8B Trigger Tokens**: This model proved sensitive to more functional and modal terms: *need, rational,* and *acceptable*. The most dramatic shifts occurred when using the Anxious wrapper, particularly affecting the Commonsense and Virtue domains.

C. Comparative Fragility Analysis

The divergence between a 10.0% and a 40.0% inconsistency rate highlights two different architectures of failure. OpenAI acts as a relatively stable anchor, where shifts are rare but predictable based on specific authoritative triggers. Conversely, Llama-3-8B manifests a "Liquid Morality," where the ethical judgment is in a state of constant flux depending on the emotional descriptors used in the system prompt. The 50.0% discrepancy in the Justice category is particularly alarming, as it suggests that principles of fairness are the most susceptible to being distorted by the stylistic "noise" of personality induction.

VI. CONCLUSION

A. Summary of Findings

This research has demonstrated that moral alignment in current Large Language Models is characterized by a "Liquid Morality," where ethical verdicts are highly susceptible to the linguistic context of personality wrappers. Our data reveals a critical disparity in resilience: while OpenAI maintains a more stable profile with a **10.0% inconsistency rate**, Llama-3-8B exhibits significant volatility, changing its judgment in **40.0% of cases**, with a peak of **50.0% in the Justice category**. Furthermore, we documented a systemic asymmetry in ethical detection, notably OpenAI's **0.0% accuracy** in identifying immoral deontological violations and moral virtuous acts. The token masking analysis confirmed that specific "trigger words" like *authoritative* or *anxious* can shift conviction by up to **30.0%**, proving that stylistic noise can effectively override internal moral logic.

B. Implications for AI Safety and Alignment

The implications for AI safety are profound. The high sensitivity to personality wrappers suggests that existing alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF), may only produce a "surface-level" compliance rather than a robust internal ethical framework. If an autonomous

system can be manipulated into justifying unethical behavior simply through a shift in prompt tone (sycophancy), its deployment in high-stakes environments such as legal reasoning or medical ethics—poses a substantial risk. The observed **Algorithmic Rigorism** and categorical blindness indicate that current safety guardrails are prone to systemic failures when faced with specific linguistic personas.

C. Limitations and Future Research

While this study provides a detailed mapping of moral fragility, it is limited by its scope of 150 scenarios and two specific model architectures. Future research should expand this methodology to larger datasets and more diverse models, such as Claude 3.5 or GPT-4o, to verify the universality of these vulnerabilities. Additionally, there is an urgent need to develop "persona-invariant" alignment strategies. Investigating synthetic data training to decouple moral reasoning from stylistic cues represents a promising avenue for creating AI systems that possess a truly stable and objective ethical compass, resilient to both adversarial manipulation and tonal influence.

REFERENCES

[1] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt, "Aligning ai with shared human values," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[2] J. Wei, D. Da, A. Qiao *et al.*, "Simple synthetic data reduces sycophancy in large language models," *arXiv preprint arXiv:2308.03958*, 2023.

[3] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *arXiv preprint arXiv:2306.05685*, 2023.