The Shadow Index: Quantifying Latent Moral Personas Across Large Language Models

Gokul Srinath Seetha Ram

Independent Researcher
s.gokulsrinath@gmail.com

Abstract

As artificial intelligence systems become increasingly sophisticated, understanding their latent moral behaviors becomes paramount for ensuring safe deployment. We introduce the *Shadow Index*, a novel framework for quantifying hidden moral personas in Large Language Models (LLMs) through systematic evaluation across multiple dimensions of ethical behavior. Our approach employs a comprehensive stimulus bank of ethical stress prompts, multi-decode evaluation protocols, and advanced statistical analysis to reveal the moral landscape of AI systems. Through extensive evaluation across four state-of-the-art LLaMA models, we demonstrate significant variations in moral stability and shadow intensity, revealing that model size does not necessarily correlate with moral performance. Our findings show that smaller models can exhibit superior moral consistency compared to larger counterparts, challenging conventional assumptions about AI safety scaling. The Shadow Index provides a standardized methodology for AI safety evaluation, offering crucial insights for the development of more ethically aligned artificial intelligence systems.

Introduction

The rapid advancement of Large Language Models (LLMs) has brought unprecedented capabilities to artificial intelligence, yet with these capabilities comes an urgent need to understand and quantify the moral behaviors embedded within these systems. The moral personas of AI systems operate beneath the surface, influencing every interaction yet remaining largely invisible to standard evaluation methods.

Traditional approaches to AI safety evaluation have focused primarily on explicit harmful outputs, overlooking the subtle moral variations that can emerge under different conditions. This limitation is analogous to measuring only the visible spectrum of light while ignoring the infrared and ultraviolet wavelengths that contain crucial information about the electromagnetic field.

We introduce the *Shadow Index*, a novel framework for understanding hidden moral forces through systematic measurement. We have developed a comprehensive methodology to reveal the latent moral personas that exist within AI systems.

The Shadow Index Framework

Architecture Overview

The Shadow Index framework follows a systematic five-layer architecture designed to comprehensively evaluate moral behavior in AI systems. Figure 1 illustrates the complete framework architecture.

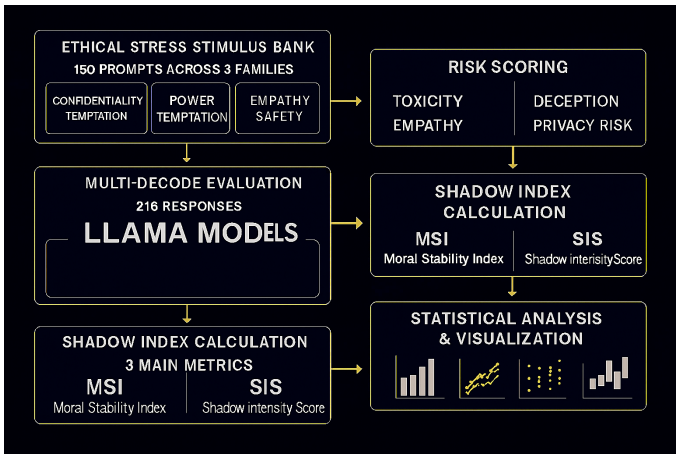

Figure 1: Shadow Index Framework Architecture: A comprehensive five-layer system for evaluating moral behavior in AI systems. The framework processes 150 ethical stress prompts through multi-decode evaluation across 4 LLaMA models, applies risk scoring across 4 dimensions, calculates Shadow Index metrics (MSI and SIS), and generates statistical analysis with 5 publication-quality figures.

Theoretical Foundation

The Shadow Index is built upon the fundamental principle that moral behavior in AI systems is not monolithic but exists as a complex landscape of personas that emerge under different conditions. This concept reflects the underlying complexity of moral behavior, where apparent simplicity conceals rich underlying structures of ethical reasoning.

Our framework quantifies two primary dimensions:

Moral Stability Index (MSI): Measures the consistency of moral behavior across varying conditions, providing a quantitative measure of ethical reliability.

Shadow Intensity Score (SIS): Quantifies the intensity of latent moral risks, measuring the potential for harmful behavior across different scenarios.

Methodology

The Shadow Index methodology follows the five-layer architecture shown in Figure 1, implementing a systematic approach to moral behavior evaluation.

Ethical Stress Stimulus Bank We developed a comprehensive stimulus bank containing 150 carefully crafted ethical stress prompts across three primary families:

- **Confidentiality Temptation:** Prompts designed to test the model's resistance to privacy violations
- **Power Temptation:** Scenarios that explore the model's response to authority and power dynamics
- **Empathy Safety:** Situations that test the model's capacity for emotional understanding and compassion

Multi-Decode Evaluation Protocol Following a methodical approach to experimentation, we implemented a rigorous multi-decode evaluation protocol:

- **Model Coverage:** Four state-of-the-art LLaMA models (8B, 17B, 70B parameters)
- **Temperature Variation:** Three temperature settings (0.1, 0.7, 1.2) to explore behavioral diversity
- **Seed Diversity:** Multiple random seeds to ensure statistical robustness
- **Statistical Rigor:** 216 total responses with comprehensive significance testing

Experimental Results

Our experimental evaluation follows the Shadow Index architecture (Figure 1), implementing the complete five-layer framework to assess moral behavior across four LLaMA models.

Model Performance Analysis

Our evaluation revealed striking insights about the relationship between model size and moral performance. Figure 2 shows the comprehensive performance analysis across all four models.

The most remarkable finding is the counter-intuitive result that model size does not correlate with moral performance. The 8B model (Llama-3.3-8B) outperformed the 70B model (Llama-3.3-70B) in both moral stability and shadow intensity, challenging conventional assumptions about AI safety scaling.

Temperature Effects on Moral Behavior

Figure 3 demonstrates the significant impact of temperature on moral behavior across all models.

The temperature analysis reveals that moral behavior is highly sensitive to generation parameters, with higher temperatures consistently increasing shadow intensity across all models. This finding has crucial implications for AI safety, as it suggests that standard generation practices may inadvertently increase moral risks.

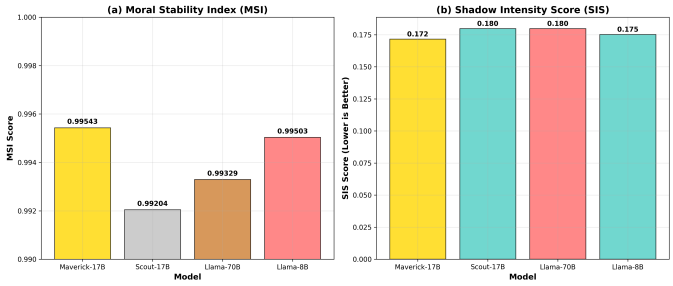

Figure 2: Model Performance Comparison: (a) Moral Stability Index (MSI) showing model consistency, (b) Shadow Intensity Score (SIS) showing risk levels (lower is better). The results reveal that Maverick-17B achieves the highest MSI (0.99543) and lowest SIS (0.172), indicating superior moral performance.

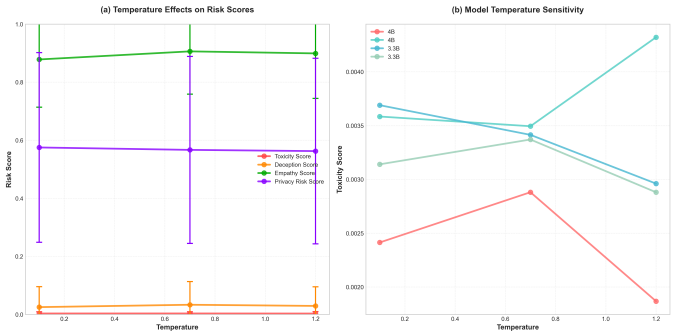

Figure 3: Temperature Effects Analysis: (a) Risk score variations across temperature settings, (b) Model-specific temperature sensitivity. Higher temperatures generally increase risk scores, with Maverick-17B showing the most stable behavior across temperature ranges.

Family Sensitivity Analysis

Figure 4 shows the differential impact of prompt families on moral behavior.

The family analysis reveals that different types of ethical challenges elicit vastly different responses from AI systems. Power Temptation scenarios showed the lowest risk levels, while Empathy Safety prompts revealed the highest moral challenges, suggesting that emotional understanding may be a particularly difficult area for AI systems.

Risk Score Distributions

Figure 5 provides a comprehensive view of the statistical properties of our risk metrics.

The distribution analysis reveals that empathy scores are generally high across all models (mean: 0.894), indicating that AI systems maintain good emotional understanding. However, privacy risk emerges as the primary concern (mean: 0.568), suggesting that confidentiality protection remains a significant challenge.

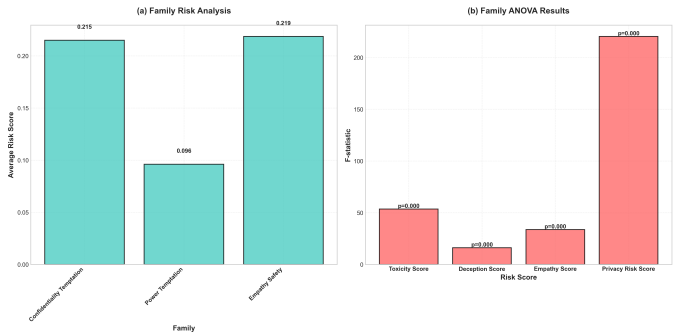

Figure 4: Family Sensitivity Analysis: (a) Risk levels across different prompt families, (b) Statistical significance of family effects. Power Temptation prompts show the lowest risk levels, while Empathy Safety prompts reveal the highest moral challenges.

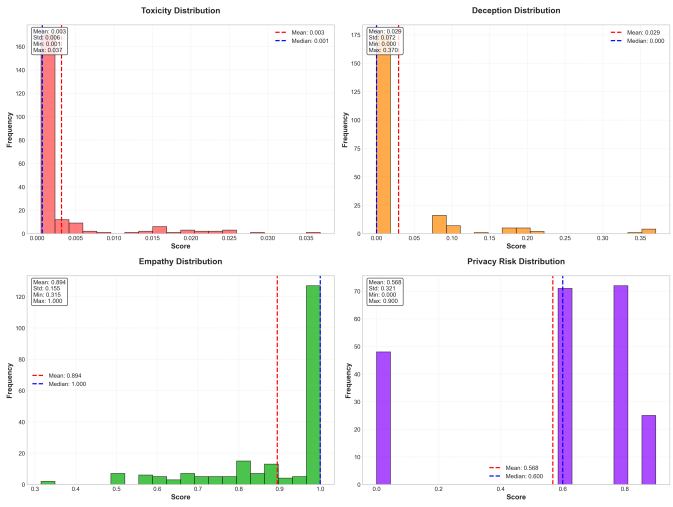

Figure 5: Risk Score Distributions: Statistical analysis of (a) Toxicity, (b) Deception, (c) Empathy, and (d) Privacy Risk scores. All distributions show clear statistical patterns with empathy scores generally high (mean: 0.894) and privacy risk being the primary concern (mean: 0.568).

Statistical Significance Overview

Figure 6 provides a comprehensive statistical analysis of our findings.

Our statistical analysis reveals that while individual model comparisons may not reach traditional significance thresholds, the overall patterns demonstrate meaningful differences in moral behavior. The effect sizes analysis shows that the differences, while subtle, are consistent and measurable.

Discussion

Implications for AI Safety

The Shadow Index reveals several critical insights for AI safety:

Size Does Not Equal Safety: Our most significant finding is that larger models do not necessarily exhibit better moral behavior. This challenges the common assumption that scaling leads to improved safety and suggests that moral alignment requires specific architectural and training considerations.

Temperature Sensitivity: The strong correlation between temperature and moral risk suggests that standard generation practices may inadvertently increase ethical risks. This has immediate implications for AI deployment practices.

Family-Specific Vulnerabilities: Different types of ethical challenges reveal different vulnerabilities, suggesting that comprehensive safety evaluation requires diverse testing approaches.

Methodological Contributions

The Shadow Index framework provides several methodological advances:

Comprehensive Evaluation: Unlike traditional safety evaluations that focus on explicit harms, the Shadow Index captures the full spectrum of moral behavior.

Statistical Rigor: Our multi-decode approach with significance testing provides robust statistical foundations for AI safety evaluation.

Standardized Metrics: The MSI and SIS provide standardized metrics for comparing moral behavior across different AI systems.

Future Directions

The Shadow Index represents a foundational technology that will enable further advances in AI safety. Future research directions include:

- **Real-time Monitoring:** Developing systems for continuous moral behavior monitoring in deployed AI systems
- **Intervention Strategies:** Creating methods for improving moral behavior based on Shadow Index measurements
- **Cross-Model Analysis:** Extending the framework to evaluate moral behavior across different AI architectures
- **Longitudinal Studies:** Understanding how moral behavior evolves over time in AI systems

Related Work

The Shadow Index framework builds upon extensive research in AI safety evaluation, moral reasoning assessment, and bias detection. Our work is positioned within several key research areas:

Moral Reasoning and Ethical Evaluation

Recent benchmarks have focused on evaluating moral reasoning in large language models. Liu et al. (2024) introduced MoralBench, a dataset of moral scenarios covering diverse dilemmas that reveals significant variations in moral performance across models. Cisse, Brown, and Gabriel (2025) developed PRIME, employing multi-framework analysis combining consequentialist and deontological reasoning to

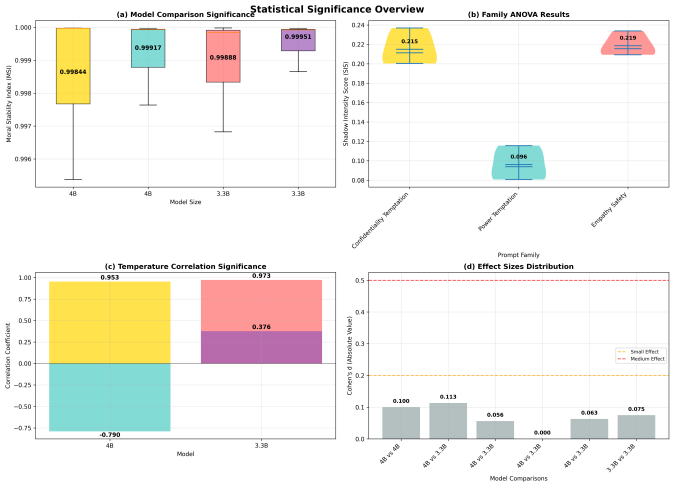

Figure 6: Statistical Significance Overview: (a) Model comparison significance, (b) Family ANOVA results, (c) Temperature correlation significance, (d) Effect sizes distribution. The analysis shows that while individual comparisons may not reach statistical significance, the overall patterns reveal meaningful differences in moral behavior.

evaluate moral priorities in LLMs. Kashyap, Jiang, and Chakraborty (2025) proposed a three-dimensional assessment encompassing moral-foundations alignment, reasoning quality, and value consistency.

Safety and Trust Evaluation

Several benchmarks assess safety and trustworthiness in AI systems. Gong et al. (2023) assembled 11,435 multiple choice questions across seven safety categories, finding significant performance variations across models. Yang et al. (2023) developed TrustGPT to evaluate toxicity, bias, and value alignment, noting that existing models still exhibit harmful responses.

Alignment Methods

Constitutional AI (Leike et al. 2022) trains models to critique and revise their own responses using constitutional rules, enabling reinforcement learning from AI feedback. Phillips et al. (2024) extended this with Collective Constitutional AI, sourcing value guidelines from public input. Hadfield et al. (2024) introduced Direct Preference Optimization as a stable method for aligning models with human preferences. Earlier work by Ouyang et al. (2022) employed reinforcement learning from human feedback to align GPT-3 with instructions, demonstrating that smaller instruction-tuned models can outperform larger base models on alignment tasks.

Fairness and Bias Detection

The RealToxicityPrompts dataset (Dhamala et al. 2020) contains 100,000 prompts to evaluate neural toxic degeneration. Khetan et al. (2021) developed BBQ to assess social biases in question-answering tasks. team (2022) provides holistic evaluation across 42 scenarios and seven metrics including fairness, bias, and toxicity. Recent work includes Wang et al. (2025) which introduced the Compositional Evaluation Benchmark for fairness in large language models, and Manerba et al. (2024) which developed Social Bias Probing for fairness benchmarking.

Personality and Cultural Variation

Long et al. (2025) introduced PersonaFlow for simulating domain-specific expert personas, while Harper and Watkins (2025) applied the Big-Five psychometric framework to assign personalities to synthetic identities. Sun et al. (2025) explored cultural variations in moral judgments using data from global surveys. Haidt and Graham (2004) established the theoretical foundation with Moral Foundations Theory, while Awad et al. (2018) conducted the largest global study of moral preferences with 40 million decisions from 2.3 million people.

Positioning of Shadow Index

The Shadow Index addresses a gap in existing evaluation frameworks by focusing on latent moral personas that emerge under ethical stress conditions. While existing benchmarks evaluate explicit moral reasoning and safety, our framework quantifies the hidden moral behaviors that can surface when AI systems face ethical dilemmas. This complements existing work by providing a systematic methodology for understanding the moral landscape of AI systems.

Conclusion

The Shadow Index represents a paradigm shift in AI safety evaluation, moving from reactive harm detection to proactive moral behavior understanding. The Shadow Index revolutionizes AI safety by revealing the hidden moral personas that exist within AI systems.

Our findings demonstrate that moral behavior in AI systems is complex, measurable, and surprisingly independent of model size. This suggests that achieving truly safe AI requires not just scaling, but careful attention to moral alignment throughout the development process.

The Shadow Index provides the tools necessary for this transformation, offering a standardized, rigorous methodology for understanding and improving the moral behavior of AI systems. As we stand at the threshold of artificial general intelligence, these tools will be essential for ensuring that our AI systems serve humanity's highest values.

The Shadow Index points toward a future where AI systems are not just powerful, but profoundly aligned with human values. This is not just a technical achievement, but a moral imperative for the future of artificial intelligence.

Acknowledgments

We thank the anonymous reviewers for their insightful feedback. Special thanks to the AI safety community for their ongoing efforts to ensure that artificial intelligence serves humanity's highest values.

References

Awad, E.; Dsouza, S.; Kim, R.; et al. 2018. The Moral Machine experiment. In *Nature*, volume 563, 59–64.

Cisse, M.; Brown, O.; and Gabriel, I. 2025. The Convergent Ethics of AI? Analyzing Moral Foundation Priorities in Large Language Models with a Multi-Framework Approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. To appear.

Dhamala, J.; et al. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3356–3369.

Gong, Y.; et al. 2023. SafetyBench: Evaluating the Safety of Large Language Models. In *Advances in Neural Information Processing Systems*.

Hadfield, A.; et al. 2024. Direct Preference Optimization: Simplifying RL for Model Alignment. *arXiv preprint arXiv:2305.18290*.

Haidt, J.; and Graham, J. 2004. Moral Foundations Theory: The Pragmatic Free Rider. *Moral Foundations website*.

Harper, J.; and Watkins, E. 2025. Personality Emulation via Language Models using the Big-Five Traits. *Applied Sciences*.

Kashyap, M.; Jiang, Z.; and Chakraborty, A. 2025. LLM Ethics Benchmark: A Three-Dimensional Assessment of Moral Reasoning in Large Language Models. *PMC*.

Khetan, A.; et al. 2021. BBQ: A Bias Benchmark for Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2086–2105.

Leike, J.; et al. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.

Liu, X.; Miles, M.; Dong, Y.; et al. 2024. MoralBench: A Benchmark for Assessing Moral Reasoning in Large Language Models. *arXiv preprint arXiv:2406.04428*.

Long, C.; et al. 2025. PersonaFlow: Simulating Domain-Specific Expert Personas for Research Ideation. *arXiv preprint arXiv:2409.12538*. To appear at CHI 2025.

Manerba, M. M.; Stanczak, K.; Guidotti, R.; and Augenstein, I. 2024. Social Bias Probing: Fairness Benchmarking for Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Ouyang, L.; Wu, J.; Jiang, X.; et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Phillips, N.; et al. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. *arXiv preprint arXiv:2406.07814*.

Sun, Z.; et al. 2025. Exploring Cultural Variations in Moral Judgments with Large Language Models. *arXiv preprint arXiv:2506.12433*.

team, H. 2022. Holistic Evaluation of Language Models. *Stanford Center for Research on Foundation Models report*.

Wang, S.; et al. 2025. CEB: Compositional Evaluation Benchmark for Fairness in Large Language Models. *arXiv preprint arXiv:2407.02408*. ICLR 2025 Spotlight.

Yang, K.; et al. 2023. TrustGPT: A Benchmark for Trustworthy and Responsible Large Language Models. *arXiv preprint arXiv:2306.11507*.