Human-AI Moral Judgment Congruence on Real-World Scenarios: A Cross-Lingual Analysis

Nan Li and Bo Kang and Tijl De Bie
IDLab, Department of Electronics and Information Systems
Ghent University, Belgium

Abstract

As Large Language Models (LLMs) are deployed in every aspect of our lives, understanding how they reason about moral issues becomes critical for AI safety. We investigate this using a dataset we curated from Reddit's r/AmItheAsshole, comprising real-world moral dilemmas with crowd-sourced verdicts. Through experiments on five state-of-the-art LLMs across 847 posts, we find a significant and systematic divergence where LLMs are more lenient than humans. Moreover, we find that translating the posts into another language changes LLMs' verdicts, indicating their judgments lack cross-lingual stability.

1 Introduction

As LLMs become ubiquitous across applications, understanding their moral reasoning becomes critical for AI safety and for predicting their congruence with human values. Current benchmarks for moral reasoning use simple problems. These problems are not like the complex moral situations in real life. Also, most studies only test in English. This leaves two important questions open: how congruent LLM judgments are with human consensus in daily personal conflicts, and if their moral reasoning is consistent across different languages.

To address these limitations, we curated a benchmark from Reddit's r/AmItheAsshole (AITA), a dataset of everyday moral conflicts with crowd-sourced verdicts. We use these verdicts as a benchmark for majority human opinion rather than objective moral truth. We investigate how five state-of-the-art LLMs judge these scenarios compared to this human consensus, and how their performance changes when scenarios are presented in English versus Chinese.

2 Related Work and Motivation

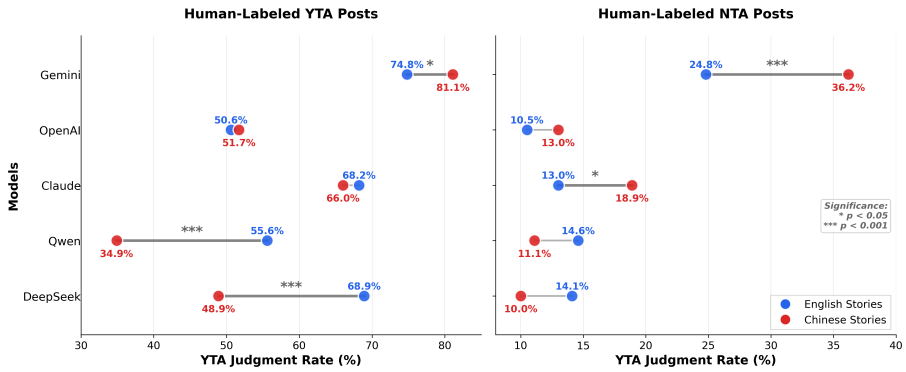
Moral Reasoning in LLMs: Recent work has increasingly focused on evaluating moral reasoning capabilities of LLMs, with a particular emphasis on alignment with human judgment. Forbes et al. (2020) introduced Social-Chem-101, using AITA posts as a testbed for social norm reasoning. Subsequent studies revealed consistent biases, with models showing systematic leniency toward morally questionable behavior, leading to poor alignment with human consensus (Malmqvist, 2024; Pratik S. Sachdeva and van Nuenen, 2025). However, these studies primarily focus on accuracy metrics rather than understanding the underlying causes of human-AI disagreement, and they only test in English.

Cross-Lingual Consistency of AI Judgment: The challenge of maintaining consistent AI behavior across languages has gained attention as models are deployed globally. While often framed as a problem of "cross-lingual alignment," studies using benchmarks like UNIMORAL (Shivani Kumar and David Jurgens, 2025) and CMoralEval (Linhao Yu et al., 2024) reveal significant variation in moral judgments across languages, echoing findings from multilingual ethical reasoning tasks (Utkarsh Agarwal et al., 2024). Large-scale audits confirm culturally rooted value differences in LLM outputs (Aksoy, 2024; Papadopoulou et al., 2024), while multilingual studies show divergent moral preferences across 100+ languages (Jin et al., 2024). But these studies use simplified moral dilemmas which are not like the complex moral situations in real life, and while reporting the statistics, the underlying reasons are not investigated.

3 Tentative Methodology

Dataset: We curated a dataset of 847 top-rated posts from Reddit's r/AmItheAsshole (AITA), collected in August 2025 and spanning the years 2018-2025. The final human verdict for each post was determined by the official moderator-assigned "flair," representing the community consensus. We cre-

Figure 1: YTA (You're the Asshole) judgment rates across five LLMs for identical moral scenarios presented in English versus Chinese. Each point represents the percentage of posts judged as YTA in each language, with lines connecting the same model's performance across languages.



ated a binary task by labeling posts with the "Asshole" flair as YTA and combining "Not the A-hole" and "No A-holes here" flairs for the NTA category. To ensure clear judgments, ambiguous or contested cases were filtered out. The final dataset contains 454 YTA and 393 NTA posts, with an average length of approximately 430 words.

Models: We evaluated five state-of-the-art LLMs: DeepSeek R1, Qwen 3, Claude 4.0 Sonnet, OpenAI o4-mini, and Gemini 2.5 Flash, chosen for their strong reasoning capabilities and multilingual support.

Prompting Strategy: For each post, we used a neutral prompt that framed the task with the direct instruction: "You are an impartial evaluator assessing a personal conflict scenario. Your task is to read the situation below and determine whether the person acted like an asshole." To ensure consistent output, the prompt further required the model to return a single JSON object with two keys: `judgment`, containing either "Y" (for YTA) or "N" (for NTA), and `explanation`, containing a brief justification.

Translation: All posts were translated to Chinese using DeepSeek R1 (temperature=0.0) with prompts designed to maintain online post style while producing truthful and natural Chinese.

Semantic Analysis: We developed heuristic keyword lists to quantify moral frameworks (deontological vs. consequentialist), cultural orientations, and reasoning styles. This allows us to trace how translation changes the underlying moral reasoning patterns that drive verdict changes.

4 Early Results

Substantial Leniency Bias: Our primary finding is a significant and systematic divergence where LLMs are more lenient than humans. As shown in Table 1, individual models and the majority vote consistently show lower agreement on posts humans labeled as YTA compared to those labeled NTA. This leniency is confirmed in Table 2. The LLM majority vote differed from the human verdict on YTA posts at a high rate (36.0%), absolving a party that the human majority found to be at fault. In contrast, the rate of divergence for NTA posts was much lower (9.9%). A McNemar's test on these discordant pairs (155 vs. 38) confirms this asymmetry is a highly significant systematic bias towards leniency ($\chi^2(1, N = 193) = 70.9, p < .001$). Our preliminary semantic analysis suggests this bias stems from models over-emphasizing practical justifications in their reasoning compared to human users.

Table 1: Model Agreement with Human Consensus.

Model	YTA Posts	NTA Posts
DeepSeek R1	69.2%	86.3%
OpenAI o4-mini	50.9%	89.3%
Gemini 2.5 Flash	73.8%	74.8%
Qwen 3	55.9%	84.7%
Claude 4.0 Sonnet	68.5%	87.0%
Majority Vote	**60.8%**	**87.5%**

Table 2: Confusion Matrix: LLM Majority Vote vs. Human Verdict (ties are excluded).

	LLM Majority Verdict		
Human Verdict	**YTA**	**NTA**	**Total**
YTA	276 (64.0%)	155 (36.0%)	431
NTA	38 (9.9%)	344 (90.1%)	382
Total	314	499	813

Translation Significantly Changes Verdicts: Our next finding is that translation dramatically alters

model judgments on identical moral scenarios. As shown in Figure 1, some models become significantly more lenient when judging Chinese translations compared to English originals. For YTA posts, DeepSeek and Qwen show the most dramatic shifts, with YTA rates dropping by over 20%, meaning they excuse behavior in Chinese that they would condemn in English. The effect was model-dependent, with Claude showing stability while other models show varying sensitivity to language.

5 Ongoing Work

This ongoing research has several directions we plan to address in the next iteration. **Understanding language effects:** The mechanism behind why language changes model verdicts remains unclear. We plan to analyze the specific content features of posts where models change their judgments to identify the underlying causes. **Improving semantic analysis:** Our current heuristic keyword matching may miss nuanced cultural concepts and context-dependent meanings. We are developing more sophisticated methods using embedding-based approaches and LLM-as-judge techniques to better capture subtle linguistic and cultural variations.

Limitations

We acknowledge several limitations that provide clear directions for future research.

Depth of Analysis: Our analysis is primarily descriptive, identifying a leniency bias and cross-lingual shifts without providing a deep causal explanation for these phenomena. The work does not include a systematic error analysis or a thematic breakdown of the dilemmas, which is our next step.

Potential Translation Confound: Our use of DeepSeek R1 for both translating the posts and for evaluation introduces a potential experimental confound. To isolate the impact of language, a future study could employ a high-quality, third-party translation model.

Dataset Scope: The generalizability of our findings is constrained by the dataset's scope. The r/AmItheAsshole community is culturally specific, primarily representing Western perspectives, and the results may not extend to other cultural contexts.

Prompt Robustness: This study utilized a single, fixed prompt to ensure experimental consistency. However, LLMs can be sensitive to variations in prompt phrasing. A valuable extension of this work would be to test for prompt robustness by using a set of semantically equivalent but lexically different prompts.

Acknowledgments

The research leading to these results has received funding from the Special Research Fund (BOF) of Ghent University (BOF20/IBF/117), from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme, from the FWO (project no. G0F9816N, 3G042220, G073924N). Funded/Co-funded by the European Union (ERC, VIGILIA, 101142229). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. For the purpose of Open Access the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Meltem Aksoy. 2024. Whose morality do they speak? unraveling cultural bias in multilingual language models. *arXiv preprint arXiv:2412.18863*.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, and 1 others. 2024. Language model alignment in multilingual trolley problems. *arXiv preprint arXiv:2407.02273*.

Linhao Yu, Yongqi Leng, Yufei Huang, and 1 others. 2024. CMoralEval: A moral evaluation benchmark for chinese large language models. In *Findings of ACL*.

Lars Malmqvist. 2024. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*.

Evi Papadopoulou, Hadi Mohammadi, and Ayoub Bagheri. 2024. Large language models as mirrors of societal moral standards. *arXiv preprint arXiv:2412.00956*.

Pratik S. Sachdeva and Tom van Nuenen. 2025. Normative evaluation of large language models with everyday moral dilemmas. *arXiv preprint arXiv:2501.18081*.

Shivani Kumar and David Jurgens. 2025. Are rules meant to be broken? understanding multilingual moral reasoning with UNIMORAL. In *ACL*.

Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. Ethical reasoning and moral value alignment of llms depend on the language we prompt them in. In *Proceedings of LREC-COLING*.