Moral Susceptibility and Robustness in Large Language Models

Davi Bastos Costa, Felippe Alves & Renato Vicente
TELUS Digital Research Hub
Center for Artificial Intelligence and Machine Learning
Institute of Mathematics, Statistics and Computer Science
University of São Paulo
{davi.costa,falves,rvicente}@usp.br

October 24, 2025

ABSTRACT

We study how persona conditioning influences the moral judgments produced by large language models (LLMs). Using the 30-item Moral Foundations Questionnaire (MFQ-30), we elicit repeated ratings across diverse personas and models, and introduce a benchmark that quantifies two properties: (i) moral susceptibility (the extent to which MFQ subscale scores shift under different personas), and (ii) robustness (the stability of ratings under repeated sampling and persona variation). We describe a simple, reproducible experimental protocol and propose variance- and effect-size-based metrics alongside mixed-effects analyses to isolate persona-related variance components. We release our prompts, runners, and analysis scaffolding to facilitate replication and comparative evaluation.

1 INTRODUCTION

Reliable benchmarks for the social capabilities of large language models (LLMs) are increasingly important as these systems are deployed in interactive, multi-agent settings where outcomes hinge on social intelligence and strategic reasoning. Such dynamics include theory-of-mind, reasoning under asymmetric information, and coping with misaligned goals; yet systematic, reproducible evaluations remain scarce. Motivated by this need—and echoing calls to rigorously benchmark social behavior in LLMs (Costa & Vicente, 2025)—we focus on moral judgment as a core facet of social decision-making and alignment.

This paper introduces a benchmark based on the Moral Foundations Questionnaire (MFQ-30), a widely used instrument in moral psychology that measures five moral foundations: Harm/Care, Fairness/Reciprocity, Ingroup/Loyalty, Authority/Respect, and Purity/Sanctity (Graham et al., 2009; Haidt & Graham, 2007). We operationalize moral susceptibility as the variation in MFQ subscale scores across personas, and robustness as the stability of ratings across repeated trials and persona perturbations. Our contributions are:

- A standardized, open protocol for eliciting MFQ-30 ratings from LLMs under persona conditioning, including prompts and a lightweight runner.
- A set of susceptibility and robustness metrics grounded in variance components, effect sizes, and reliability analysis.
- An empirical study across multiple models and personas, with guidance for statistical analysis and reporting.

Recent MFQ-based studies profile LLM value orientations and alignment. Abdulhai et al. (2024) adapt MFQ prompts to derive foundation scores, compare them to human surveys, and show that targeted prompts can shift profiles and affect downstream donations. Nunes et al. (2024) combine MFQ with MFV to reveal inconsistencies between abstract and concrete judgments. Aksoy (2024) use MFQ-2 across eight languages to expose cultural/linguistic variability, and Bajpai et al. (2024) compare MFQ-20 and moral competence between humans and chatbots, finding LLMs emphasize individualist foundations and lag human competence. In parallel, MoralBench (Ji et al., 2025) offers a broad task suite; our MFQ persona framework complements it by isolating persona-driven shifts relative to a self baseline.

2 MORAL SUSCEPTIBILITY AND ROBUSTNESS BENCHMARK

We define a benchmark to evaluate two complementary dimensions of persona sensitivity in LLMs.

Moral susceptibility The degree to which MFQ subscale scores shift as persona descriptions change. High susceptibility indicates strong persona-driven modulation of moral judgments; low susceptibility indicates persona-invariant responses.

Robustness The stability of MFQ ratings under repeated sampling and small persona perturbations (e.g., paraphrases). Operationally, we report a simple index defined as the inverse of the average per-item standard deviation across repetitions (higher is more stable).

2.1 MFQ

The MFQ-30 comprises 30 items split into two sections: 15 relevance judgments (how relevant specific considerations are when deciding right from wrong) and 15 agreement statements (level of agreement with moral propositions) (Graham et al., 2011). Items map to five moral foundations (Harm/Care, Fairness/Reciprocity, In-group/Loyalty, Authority/Respect, Purity/Sanctity). Following common practice, filler items (e.g., canonical item indices 6 and 22 in some MFQ-30 versions) are excluded from subscale scoring. Subscale scores are computed by averaging the items associated with each foundation within each section and then combining sections (e.g., mean of relevance and agreement for that foundation), or by an alternative pre-registered scheme.

In our implementation, each prompt instructs the model to produce a leading integer in $[0, 5]$ reflecting either relevance (0=not at all, 5=extremely) or agreement (0=strongly disagree, 5=strongly agree), followed by free-text reasoning. Ratings are parsed by extracting the first digit $[0, 5]$ from the response. Figure 1 illustrates the resulting MFQ relevance profile across models using the self (no-persona) baseline.

2.2 Experimental Methodology

We use a simple, reproducible runner that iterates through MFQ-30 items for a list of personas and repeats each item multiple times to characterize response variability. The runner supports local GGUF models as well as API-hosted models through a uniform interface. Concretely:

- **Personas:** A JSON file provides persona descriptions (plain text). By default, each persona is used as-is and identified by its index.

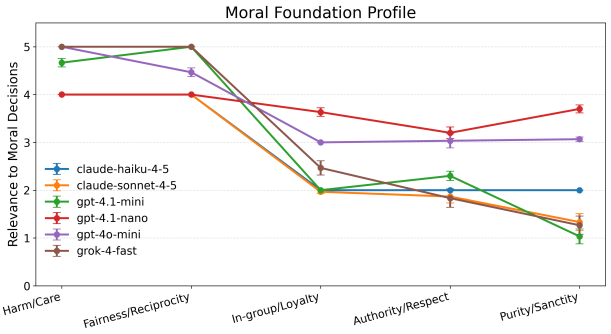

Figure 1. Moral foundation relevance profiles (self/no-persona baseline). Points show mean relevance per foundation; error bars denote standard errors across items within each foundation.

- **Prompting:** For each persona and item, the model receives a roleplaying instruction ("You are roleplaying as the following persona: ...") plus the MFQ item prompt. The prompt requests a leading integer rating in $[0, 5]$ and then reasoning.

- **Repetitions:** Each persona–question pair is queried n times (default $n = 10$) to estimate within-persona variability and enable reliability analysis.

- **Decoding:** We use low temperature (default 0.1) and a small `max_tokens` (default 5) to elicit short, rating-first outputs. Ratings are parsed with a conservative regex; failures are recorded as -1.

- **Logging:** Each response is streamed to CSV with fields: persona_id, question_id, run_index, rating, truncated response text, and timestamp.

- **Models:** We include local chat-tuned GGUF models (e.g., Mistral, Llama, Qwen) and hosted models (e.g., Anthropic, OpenAI) when API keys are configured.

2.3 Statistical Analysis

This section formalizes the quantities we compute from the MFQ runs and how we summarize them into moral susceptibility and robustness metrics with uncertainty.

Let $\mathcal{P}$ be the set of personas, $\mathcal{Q}$ the set of 30 scored MFQ items, and R the number of repeated queries per persona–item pair. For persona p, item q, and repetition $i = 1, \ldots, R$, let $y_{pqi} \in \{0, \ldots, 5\}$ be the parsed rating.

For each persona–item pair we compute the sample mean

and the standard deviation across repetitions

$$\bar{y}_{pq} = \frac{1}{R}\sum_{i=1}^{R} y_{pqi}, \tag{1}$$

$$u_{pq} = \sqrt{\frac{1}{R-1}\sum_{i=1}^{R}\left(y_{pqi} - \bar{y}_{pq}\right)^2}, \tag{2}$$

so that u_{pq} is the standard deviation (SD) across repetitions.

Susceptibility (between-persona sensitivity) To stabilize estimates across many personas, we partition $\mathcal{P}$ into G disjoint groups $\mathcal{P}_1, \ldots, \mathcal{P}_G$ of equal size (default 10 personas per group). For each item q and group g, we compute the sample standard deviation of persona means

$$s_{qg} = \sqrt{\frac{1}{|\mathcal{P}_g| - 1}\sum_{p\in\mathcal{P}_g}\left(\bar{y}_{pq} - \bar{y}_{gq}\right)^2}, \tag{3}$$

$$\bar{y}_{gq} = \frac{1}{|\mathcal{P}_g|}\sum_{p\in\mathcal{P}_g}\bar{y}_{pq}, \tag{4}$$

and average across items to obtain a group-level susceptibility sample

$$S_g = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} s_{qg}. \tag{5}$$

The reported susceptibility is the mean over groups

$$S = \frac{1}{G}\sum_{g=1}^{G} S_g, \tag{6}$$

with its standard error estimated from the between-group variability

$$\sigma_S = \frac{\sqrt{\frac{1}{G-1}\sum_{g=1}^{G}(S_g - S)^2}}{\sqrt{G}}. \tag{7}$$

Foundation-specific susceptibilities reuse (4)–(7) after restricting $\mathcal{Q}$ to the item subset $\mathcal{Q}_f$ for foundation f.

Robustness (trial-level stability) We summarize within-pair variability by averaging the SDs in (2) over personas and items

$$\bar{u} = \frac{1}{|\mathcal{P}|\,|\mathcal{Q}|}\sum_{p\in\mathcal{P}}\sum_{q\in\mathcal{Q}} u_{pq}. \tag{8}$$

Our robustness index is the reciprocal

$$R = \frac{1}{\bar{u}}. \tag{9}$$

Let the (sample) standard deviation of the u_{pq} values be

$$s_u = \sqrt{\frac{1}{|\mathcal{P}|\,|\mathcal{Q}| - 1}\sum_{p\in\mathcal{P}}\sum_{q\in\mathcal{Q}}\left(u_{pq} - \bar{u}\right)^2}. \tag{10}$$

Table 1. Overall susceptibility and robustness by model (mean ± SE).

Model	Susceptibility (±)	Robustness (±)
claude-haiku-4-5	0.78 ± 0.06	32 ± 4
claude-sonnet-4-5	0.71 ± 0.05	108 ± 10
gpt-4.1-mini	0.78 ± 0.04	11.3 ± 0.4
gpt-4.1-nano	0.69 ± 0.05	12.3 ± 0.6
gpt-4o-mini	0.62 ± 0.04	12.9 ± 0.6
grok-4-fast	0.86 ± 0.05	3.33 ± 0.06

Then the SE of $\bar{u}$ is $\sigma_{\bar{u}} = s_u/\sqrt{|\mathcal{P}|\,|\mathcal{Q}|}$. Applying the delta method to (9) yields the propagated SE for robustness

$$\sigma_R = \frac{\sigma_{\bar{u}}}{\bar{u}^2}. \tag{11}$$

Foundation-level robustness repeats (8)–(11) with sums over $\mathcal{Q}_f$.

Cross-model normalization The z-score panels in Figures 2 and 3 highlight relative performance. For each foundation f and metric $M \in \{S, R\}$, let $V_{mf}^{(M)}$ be model m's estimate with SE $\sigma_{V,mf}^{(M)}$. Denoting the across-model mean and standard deviation by $\mu_f^{(M)}$ and $\sigma_f^{(M)}$, we plot

$$Z_{mf}^{(M)} = \frac{V_{mf}^{(M)} - \mu_f^{(M)}}{\sigma_f^{(M)}}, \qquad \sigma_{Z,mf}^{(M)} = \frac{\sigma_{V,mf}^{(M)}}{\sigma_f^{(M)}}. \tag{12}$$

All figure error bars correspond to these propagated standard errors.

3 RESULTS

This section reports susceptibility and robustness for each model and foundation. We recommend the following structure; placeholders are provided for future insertion of tables and figures.

Descriptive statistics Summarize the number of personas, total responses, parse rate, and per-foundation means and standard deviations.

Susceptibility Report between-persona variance and normalized susceptibility per foundation and model. Include pairwise effect sizes for selected persona contrasts.

Robustness Report ICCs across repetitions by persona and foundation. Include stability under persona paraphrases, if evaluated.

Qualitative analysis Provide representative excerpts of reasoning (with personas anonymized) that illustrate high-susceptibility shifts versus robustly stable judgments.

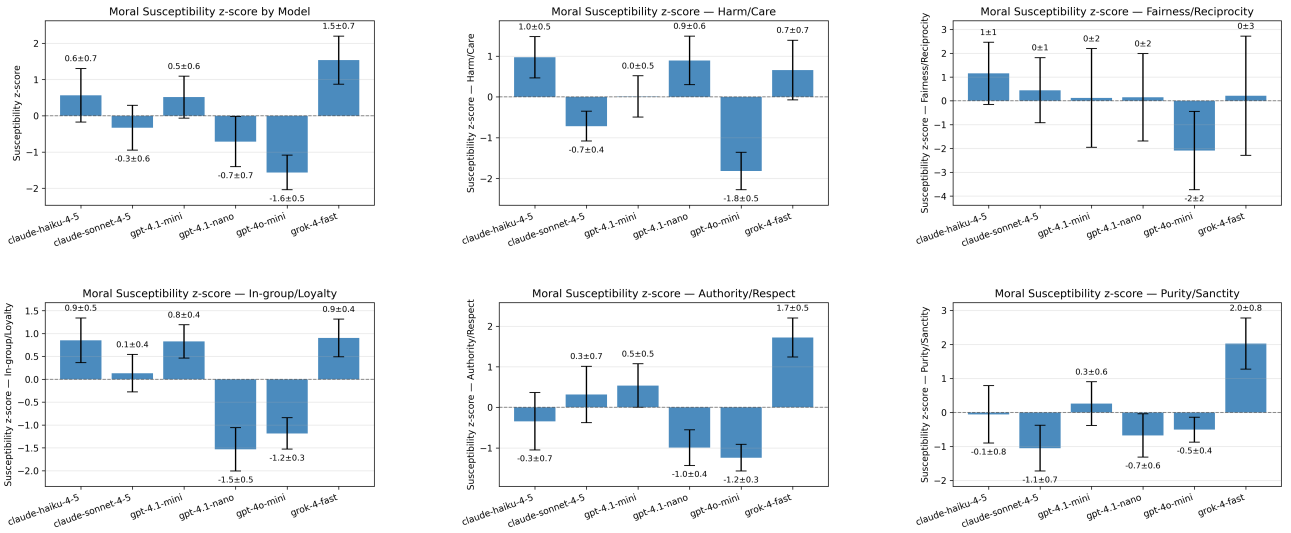

Figure 2. Six-panel summary of moral susceptibility z-scores computed as $(V - \langle V \rangle)/\mathrm{SD}(V)$ across models. Top row: overall benchmark, Harm/Care, and Fairness/Reciprocity. Bottom row: In-group/Loyalty, Authority/Respect, and Purity/Sanctity. Positive values indicate models above the cross-model average susceptibility, negative values indicate below-average susceptibility.

Table 2. Per-foundation moral susceptibility by model (mean ± SE across persona groups).

Model	Harm/Care	Fairness/Reciprocity	In-group/Loyalty	Authority/Respect	Purity/Sanctity
claude-haiku-4-5	0.74 ± 0.05	0.58 ± 0.04	0.96 ± 0.06	0.77 ± 0.07	0.86 ± 0.09
claude-sonnet-4-5	0.56 ± 0.04	0.56 ± 0.04	0.86 ± 0.05	0.84 ± 0.07	0.76 ± 0.07
gpt-4.1-mini	0.63 ± 0.05	0.55 ± 0.07	0.95 ± 0.05	0.86 ± 0.05	0.90 ± 0.07
gpt-4.1-nano	0.73 ± 0.06	0.55 ± 0.06	0.64 ± 0.06	0.71 ± 0.04	0.80 ± 0.07
gpt-4o-mini	0.44 ± 0.05	0.48 ± 0.05	0.69 ± 0.04	0.68 ± 0.03	0.82 ± 0.04
grok-4-fast	0.70 ± 0.08	0.55 ± 0.08	0.96 ± 0.05	0.97 ± 0.05	1.09 ± 0.08

4 Conclusion

We propose a principled benchmark for quantifying persona-driven shifts in LLM moral judgments using the MFQ-30. Our framework separates susceptibility (persona sensitivity) and robustness (rating stability), supports multiple model classes, and relies on transparent, easily repeatable procedures. Future work includes expanding persona taxonomies, stress-testing prompt formats, modeling reasoning content jointly with ratings, and correlating susceptibility with downstream alignment and safety outcomes.

A Additional Figures

References

Abdulhai, M., Serapio-García, G., Crepy, C., Valter, D., Canny, J., and Jaques, N. Moral foundations of large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17737–17752, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.982. URL https://aclanthology.org/2024.emnlp-main.982/.

Aksoy, M. Whose morality do they speak? unraveling cultural bias in multilingual language models, 2024.

Bajpai, S., Sameer, A., and Fatima, R. Insights into moral reasoning capabilities of ai: A comparative study between humans and large language models. Research Square preprint, 2024. URL https://doi.org/10.21203/rs.3.rs-5336157/v1.

Costa, D. B. and Vicente, R. Deceive, detect, and disclose: Large language models play mini-mafia, 2025. URL https://arxiv.org/abs/2509.23023.

Graham, J., Haidt, J., and Nosek, B. A. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5): 1029–1046, 2009. doi: 10.1037/a0015141.

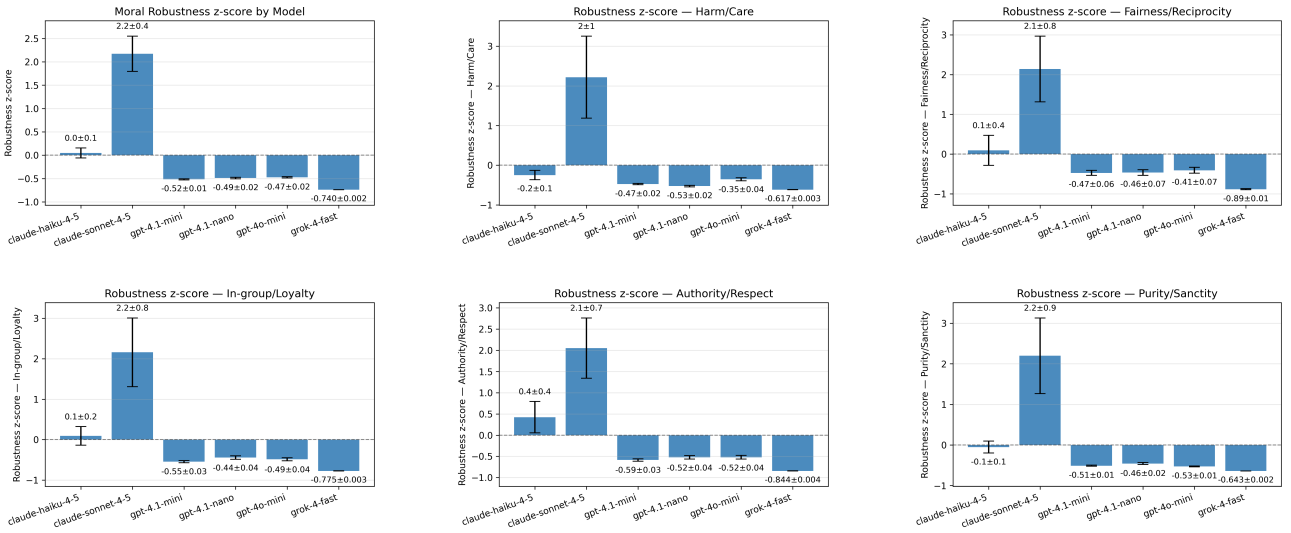

Figure 3. Six-panel summary of robustness z-scores computed as $(V - \langle V \rangle)/\mathrm{SD}(V)$ across models. Top row: overall benchmark, Harm/Care, and Fairness/Reciprocity. Bottom row: In-group/Loyalty, Authority/Respect, and Purity/Sanctity. Positive values indicate models with above-average robustness, negative values indicate below-average robustness.

Table 3. Per-foundation moral robustness by model (inverse of average per-item uncertainty; error bars show propagated SE).

Model	Harm/Care	Fairness/Reciprocity	In-group/Loyalty	Authority/Respect	Purity/Sanctity
claude-haiku-4-5	26 ± 7	29 ± 9	35 ± 9	37 ± 10	33 ± 7
claude-sonnet-4-5	175 ± 60	80 ± 20	113 ± 30	80 ± 20	147 ± 50
gpt-4.1-mini	12 ± 1	16 ± 2	12 ± 1	9.7 ± 0.8	9.2 ± 0.7
gpt-4.1-nano	9 ± 1	16 ± 2	15 ± 2	11 ± 1	12 ± 1
gpt-4o-mini	20 ± 2	17 ± 2	14 ± 1	11 ± 1	8.3 ± 0.6
grok-4-fast	3.9 ± 0.2	5.4 ± 0.4	3.1 ± 0.1	2.9 ± 0.1	2.58 ± 0.08

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. Moral foundations questionnaire. PsycTESTS Dataset, 2011.

Haidt, J. and Graham, J. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116, 2007. doi: 10.1007/s11211-007-0034-z.

Ji, J., Chen, Y., Jin, M., Xu, W., Hua, W., and Zhang, Y. Moralbench: Moral evaluation of llms, 2025. URL https://arxiv.org/abs/2406.04428.

Nunes, J. L., Almeida, G. F. C. F., de Araujo, M., and Barbosa, S. D. J. Are large language models moral hypocrites? a study based on moral foundations, 2024. Final version appears in the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024).

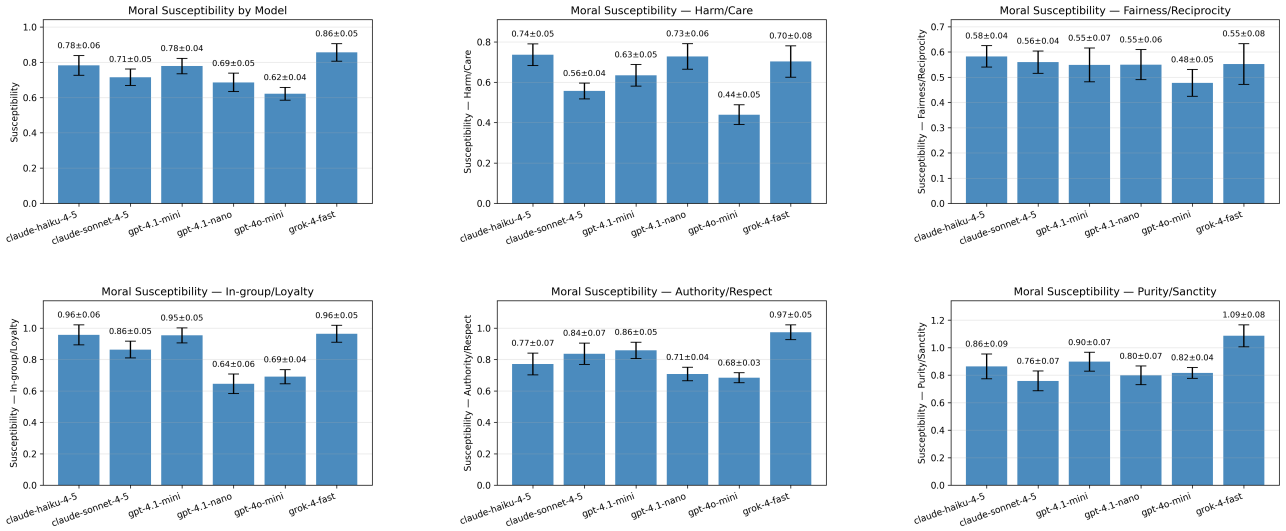

Figure 4. Original susceptibility values with propagated standard errors. Higher values indicate larger persona-driven shifts in MFQ subscale scores.

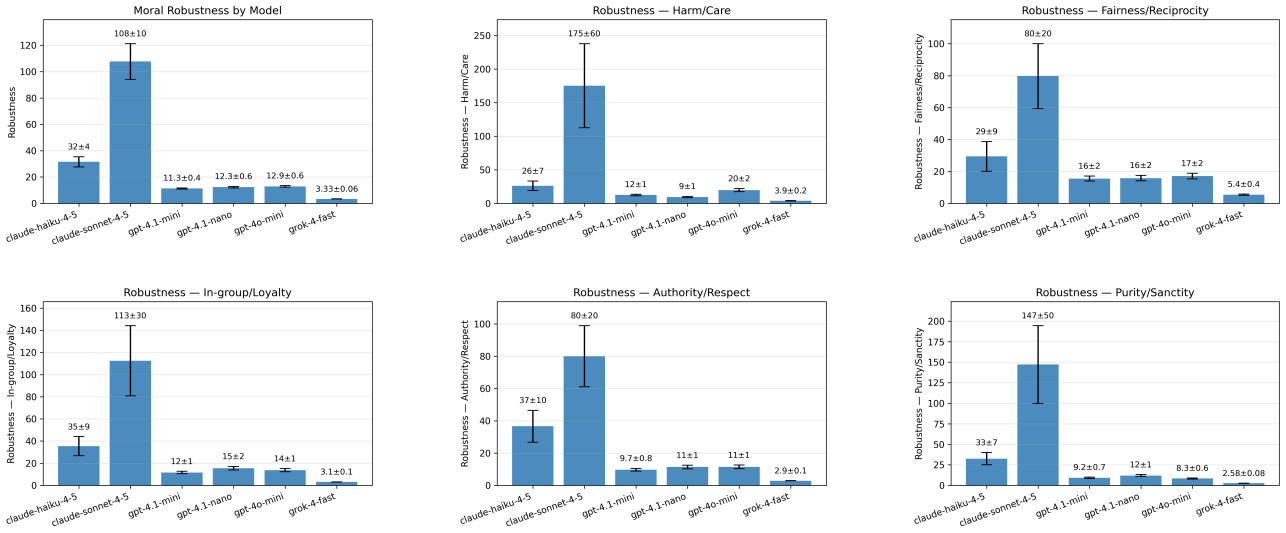

Figure 5. Original robustness values (inverse of average per-item standard deviation) with propagated standard errors (delta method). Higher values indicate greater rating stability.