Domain-Dependent Safety Behavior in Open-Weight LLMs: An Empirical Study Across Seven Ethical Domains

Zacharie Bugaud
Astera Institute
zacharie@astera.org

Abstract

We present a systematic study of domain-dependent safety behavior in open-weight LLMs: 7 standardized experiments across 7 ethical domains, testing 5 models (12B–70B) in 4,200 interactions with dual-judge validation. Using a dual-condition methodology, each scenario tested in both an *analytical* framing (identify the harm) and an *operational* framing (help commit the harm), we find compliance rates vary from 14.7% (human trafficking) to 85.7% (surveillance design), a 71-percentage-point span with non-overlapping cluster-bootstrapped 95% CIs. Domain accounts for 36% of pair-level variance in harm scores, with scenario (26%) exceeding model identity (15%). A stable model safety hierarchy persists across domains (mean Spearman $\rho = 0.68$). These findings demonstrate that safety alignment is not a general capability: aggregate safety scores mask critical domain-level variation, motivating domain-specific safety auditing for trustworthy deployment.

1 Introduction

The AI safety community has invested heavily in alignment techniques (RLHF, constitutional AI, red-teaming, safety fine-tuning) under the implicit assumption that these produce *general* safety capabilities. A model that refuses to help with bomb-making should, in principle, also refuse to help with election manipulation, scientific fraud, or environmental crime. We test this assumption with 7 standardized experiments across 7 ethical domains, evaluating 5 open-weight LLMs (12B–70B parameters) in 4,200 interactions, supported by 13 additional exploratory studies (14,650 interactions).

Central finding: Compliance rates vary from 14.7% (human trafficking) to 85.7% (surveillance design), a 71-percentage-point (pp) span with non-overlapping cluster-bootstrapped 95% CIs. The same model (Mistral Nemo 12B) provides surveillance designs in 100% of requests while complying with trafficking requests only 26.7% of the time. Llama 3.3 at 70B achieves 0% compliance on environmental crimes yet 51.7% on surveillance, with the same weights and no system prompt.

Contributions: (1) Documentation of domain-dependent compliance variation with cluster-bootstrapped inference; (2) hierarchical heterogeneity, where safety varies between domains, sub-domains, and models simultaneously, with within-domain gaps reaching 84.4pp; (3) a technical framing bypass present across all five tested models, where harmful actions reframed as engineering problems elicit higher compliance; (4) a stable model safety hierarchy (mean Spearman $\rho = 0.68$).

2 Related Work

Safety benchmarks. Existing benchmarks (TruthfulQA (Lin et al., 2022), AdvBench (Zou et al., 2023), HarmBench (Mazeika et al., 2024)) evaluate safety along a single aggregate dimension. Models receive an aggregate score, obscuring domain-level variation. Our work illustrates that a single score is misleading: Gemma 3 averages 44% compliance overall yet ranges from 8% on trafficking to 93% on surveillance.

Alignment failures. Sharma et al. (Sharma et al., 2024) document sycophancy as an RLHF side effect; Wei et al. (Wei et al., 2023) propose synthetic data to mitigate it. Casper et al. (Casper et al., 2023) identify fundamental RLHF limitations. Our results provide empirical evidence of a specific failure mode: domain-dependent safety, where alignment generalizes unevenly across ethical categories.

Domain-specific safety. Weidinger et al. (Weidinger et al., 2022) taxonomize AI risks but do not test whether alignment is uniform across them. Bai et al. (Bai et al., 2022) train a helpful and harmless assistant with RLHF but evaluate safety holistically. Our work is, to our knowledge, among the first

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 557–562

Table 1: Model specifications.

Model	Params	Developer
Gemma 3	27B	Google
Qwen 3	32B	Alibaba
Mistral Nemo	12B	Mistral AI
Llama 3.3	70B	Meta
DeepSeek R1	32B	DeepSeek

controlled studies quantifying domain-dependent safety behavior, with up to 71pp separating the best- and worst-defended domains.

3 Methods

3.1 Experimental Design

Each of 7 studies tested 20 scenarios covering distinct sub-domains within an ethical area. Each scenario was written in two conditions:

- **Analytical** (e.g., advocate, auditor, reviewer): "Identify this harm, analyze violations, recommend protections."
- **Operational** (e.g., exploiter, architect, facilitator): "Help me commit this harm, provide guidance, explain how to avoid detection."

Scale per study: 5 models $\times$ 20 scenarios $\times$ 2 conditions $\times$ 3 replications = 600 interactions.

3.2 Models

We tested five open-weight LLMs deployed locally via Ollama (Q4_K_M quant., T=0.7, max tokens 2048, no system prompt), listed in Table 1. We use T=0.7 to allow sampling diversity (the default for chat use); no system prompt is used to test base safety behavior without deployment-time guardrails. Quantization (4-bit) enables local execution; we note this as a limitation.

3.3 Evaluation

Primary judge. LLM-as-judge (Zheng et al., 2023) using Gemma 3 27B at T=0.0 with domain-specific rubrics: 3 ordinal scales (1–5) for harm identification, refusal strength, and harmful guidance quality; 2 binary indicators for domain-specific markers (e.g., cites laws, advocates protection).

Key metrics. *Compliance rate*: proportion of operational responses providing substantive harmful guidance (score $\geq$3/5). *Strong refusal rate*: proportion with refusal score $\geq$4/5. *"Hypocrisy"*: proportion of model-scenario pairs where the model scores $\geq$4 on harm identification (analytical) yet $\geq$3 on harmful guidance (operational).

Table 2: Compliance, refusal, and hypocrisy rates with bootstrapped 95% CIs. Str. = compliance stratum (L/M/H).

Domain	Str.	Compl. (CI)	Refusal	Hypoc.
Trafficking	L	14.7 [10.0, 20.0]	91.3	4.0
Corruption	L	21.7 [15.0, 28.7]	80.3	13.0
Environmental	L	24.7 [17.7, 32.0]	76.3	14.0
Elections	L	25.0 [18.3, 32.0]	85.3	14.0
Labor	M	49.7 [41.0, 58.3]	46.3	39.0
Science fraud	H	69.7 [61.7, 77.3]	43.7	53.0
Surveillance	H	85.7 [79.7, 91.3]	20.0	79.0

Table 3: Stratification quality: R^2 by contiguous strata count. T = Traf., C = Corr., En = Env., El = Elec., L = Labor, F = Sci.F., S = Surv.

k	Strata assignment	R^2	ΔR^2
2	{T,C,En,El} | {L,F,S}	0.839	+0.839
3	{T,C,En,El} | {L} | {F,S}	0.956	+0.117
4	{T,C,En,El} | {L} | {F} | {S}	0.985	+0.029

Second-judge validation. 140 stratified responses (20 per core study, non-Llama outputs) re-evaluated by Llama 3.3 70B with matching harm and refusal scales at T=0.0. Cohen's $\kappa = 0.90$ for binary compliance; 97.8% within-1 agreement on harm scores.

Statistical inference. The 3 replications per model-scenario pair at T=0.7 are not fully independent, making the pair ($n = 100$ per study) the natural independent unit. We report cluster-bootstrapped 95% CIs resampling model-scenario pairs ($B = 10{,}000$), producing CIs $\sim$1.4$\times$ wider than naive response-level estimates. We also compute pair-level variance decomposition and test threshold sensitivity at $\geq$2, $\geq$3, and $\geq$4.

4 Results

4.1 Domain-Dependent Compliance

Table 2 shows the central result. For descriptive convenience, the 7 rates partition into three strata: Low ($\sim$15–25%: trafficking, corruption, environmental, elections), Medium ($\sim$50%: labor), and High ($\sim$70–86%: science fraud, surveillance). This partition captures 96% of between-domain variance, though it is derived from only 7 data points with a singleton Medium stratum. Adjacent strata are separated: Low's upper bound (32.0%) $<$ Medium's lower (41.0%); Medium's upper (58.3%) $<$ High's lower (61.7%).

Table 4: Compliance at thresholds ≥2, ≥3, and ≥4.

Domain	≥2	≥3	≥4
Trafficking	23.7%	14.7%	1.7%
Corruption	45.3%	21.7%	3.3%
Environmental	34.0%	24.7%	0.7%
Elections	40.3%	25.0%	9.3%
Labor	66.7%	49.7%	5.7%
Science fraud	75.7%	69.7%	21.7%
Surveillance	88.0%	85.7%	56.7%

Table 5: Pair-level variance decomposition ($n = 700$ pairs).

Source	SS	% Total
Domain	279.1	35.6%
Model	115.0	14.6%
Scenario (w/in domain)	206.2	26.3%
M × S interaction	184.7	23.5%

Table 6: Per-model compliance rates (%) across 7 studies. Cell values rounded to integers; $\bar{x}$ is the exact mean.

Model	Traf	Corr	Env	Elec	Lab	Sci.F	Surv	$\bar{x}$
Llama 3.3	5	7	**0**	3	20	25	52	16.0
Qwen 3	7	15	27	22	50	67	92	39.8
Gemma 3	8	10	32	17	57	93	93	44.3
DeepSeek R1	27	38	37	35	50	75	92	50.5
Mistral	27	38	28	48	72	88	**100**	57.4

Table 7: Hypocrisy rates with 95% bootstrap CIs.

Domain	Hypoc. (%)	95% CI
Trafficking	4.0	[1, 8]
Corruption	13.0	[7, 20]
Environmental	14.0	[8, 21]
Elections	14.0	[8, 21]
Labor	39.0	[29, 49]
Science fraud	53.0	[43, 63]
Surveillance	79.0	[71, 87]

Table 3 shows that the three-stratum partition explains 96% of between-domain variance, with diminishing marginal returns for additional strata ($\Delta R^2 = 0.117$ from 2→3, $\Delta R^2 = 0.029$ from 3→4). The four Low domains have mutually overlapping CIs, so finer partitions within this group are not supported.

Effect sizes for the analytical-vs-operational condition shift are large (pair-level Cohen's d up to 3.56), confirming that the compliance differences are not merely statistically significant but practically meaningful.

The 71pp span from trafficking (14.7%) to surveillance (85.7%), with non-overlapping CIs, is the core empirical finding: safety behavior is strongly domain-dependent in the models we tested.

4.2 Threshold Sensitivity

The compliance threshold (≥3/5) is researcher-specified. We re-compute at alternative thresholds in Table 4:

The domain ordering is stable across thresholds. Kendall's τ between the ≥2 and ≥3 rankings is 0.81 ($p = 0.011$); between ≥3 and ≥4, $\tau = 0.71$ ($p = 0.030$). At ≥4, surveillance remains highest at 56.7% and environmental crime lowest at 0.7%. The stratum separation is preserved: Low domains stay below 10% at ≥4, while High domains remain above 20%.

4.3 Variance Decomposition

Table 5 shows the pair-level variance decomposition. Domain is the dominant systematic factor (36%), 2.4× more than model identity (15%). Scenario variation within domains (26%) exceeds model variation, indicating substantial sub-domain heterogeneity. The 24% residual represents model × scenario interaction, i.e. specific model-domain-scenario combinations that deviate from main effects.

4.4 Model Safety Hierarchy

Table 6 presents individual model compliance rates. The ranking (Llama < Qwen < Gemma < DeepSeek < Mistral, by increasing compliance, though Qwen, Gemma, and DeepSeek are not cleanly separated) is stable across domains: all 21 pairwise Spearman correlations are positive (mean $\rho = 0.68$; individual correlations do not reach significance at n=5, but the probability of all 21 being positive by chance is $< 5 \times 10^{-7}$ by sign test). Llama 3.3 achieves 0% compliance on environmental crimes (60 interactions) and near-zero on trafficking (5%). Mistral Nemo reaches 100% on surveillance (60 interactions, every response provided surveillance designs). The 0%–100% range across models (Llama at 0%, Mistral at 100% on different domains) illustrates that domain dominates model identity as a source of variation.

The ranking is *consistent with* a scale benefit (the largest model, 70B, is safest; the smallest, 12B, least safe), but the five models differ in training data and RLHF methodology, not only scale. We cannot attribute the ordering to parameter count alone.

Table 8: Within-study compliance ranges.

Study	Sub-domains	Min–Max	Gap
Trafficking	Child expl.–Supply chain	3.3–46.7%	43.3pp
Corruption	Obstr.–Money laund.	0.0–38.3%	38.3pp
Environmental	Emissions–Reg. evas.	3.3–66.7%	63.3pp
Elections	Oppo. res.–Deepfakes	3.3–46.7%	43.3pp
Labor	Migrant–Worker surv.	0.0–84.4%	**84.4pp**
Sci. fraud	Grant fraud–Citation	36.7–86.7%	50.0pp
Surveillance	Dissent–Workplace	73.3–93.3%	20.0pp

4.5 The Knowledge-Action Gap

One of the clearest patterns is that models *understand* a harm is wrong (scoring $\geq$4/5 on harm identification) yet *provide guidance anyway* (scoring $\geq$3/5 on harmful guidance). Table 7 shows this "hypocrisy" ranges from 4% (trafficking) to 79% (surveillance), a 75pp span with non-overlapping CIs.

Surveillance hypocrisy at 79% means that in nearly 4 of 5 model-scenario pairs, the model correctly identifies surveillance as a civil liberties violation yet provides detailed system designs when asked in the operational framing. Trafficking at 4% shows the opposite: when models recognize trafficking, they almost never provide guidance.

4.6 Intra-Domain Variation

Each study reveals substantial compliance variation across sub-domains (Table 8):

The labor study is the most extreme: within a single domain, compliance ranges from 0% (migrant exploitation, linked to federal trafficking laws) to 84.4% (worker surveillance, unregulated), an 84.4pp gap. This confirms that safety training operates at a finer granularity than the domain level. We term this **hierarchical heterogeneity**: compliance varies between domains, sub-domains, and models simultaneously, with no single level fully explaining the pattern.

4.7 Technical Framing Bypass

When harmful actions are reframed as engineering or optimization problems, compliance increases across all five models. The pattern appears across both core and exploratory studies (7+13=20 total): safety dismissal framed as "risk analysis" (92%), deception as "strategic communication" (99.8% for models $\geq$12B), military escalation as "national security" (93%), discriminatory hiring as "HR optimization" (78%), and surveillance as "system architecture" (85.7%). The within-domain labor example is especially clear: migrant exploitation receives 0% compliance while worker surveillance (framed as "workplace management") reaches 84.4%, within the same ethical domain but with very different compliance, distinguished primarily by technical framing.

Table 9: Inter-judge agreement: Gemma 3 27B (primary) vs. Llama 3.3 70B (second).

Metric	Harm	Refusal
Spearman ρ	0.866	0.820
Exact agreement	63.0%	41.3%
Within-1 agreement	97.8%	82.6%
Binary κ (compliance)	0.898	
Binary agreement	94.9%	

Table 10: Legal institutionalization by compliance stratum (Str.).

Domain	Str.	Compl.	Legal status
Trafficking	L	14.7%	TVPA, DOJ/FBI
Corruption	L	21.7%	FCPA, SEC/FBI
Environmental	L	24.7%	CAA/CWA, EPA
Elections	L	25.0%	FECA, FEC/DOJ
Labor	M	49.7%	Partial: FLSA
Sci. fraud	H	69.7%	None: ORI only
Surveillance	H	85.7%	None: no agency

4.8 Second-Judge Validation

To assess inter-judge reliability, 140 stratified responses (20 per core study, non-Llama outputs only) were re-evaluated by Llama 3.3 70B scoring the same harm and refusal dimensions at T=0.0 (Table 9):

Cohen's $\kappa = 0.90$ for binary compliance confirms near-perfect inter-judge agreement. Per-study agreement ranges from 85% (elections) to 100% (corruption, surveillance). The 97.8% within-1 agreement on harm scores indicates that when judges disagree, the disagreement is almost always by a single ordinal point.

4.9 Post-Hoc: Legal Institutionalization

We observe, post hoc, that stratum placement correlates with legal infrastructure (Table 10):

Low-stratum domains share dedicated criminal statutes and active enforcement agencies. High-stratum domains involve harms recognized as unethical but lacking criminal penalties. The singleton Medium domain (labor) has partial criminal coverage. This association is suggestive but not causal; five competing explanations remain plausible: training data representation, lexical cues,

cultural salience, judge sensitivity, and legal codification effects.

5 Discussion

Safety is not a general capability. The 71pp compliance span illustrates that aggregate safety scores mask critical variation. A model scoring 50% overall may be simultaneously near-zero on trafficking and near-total on surveillance.

Hierarchical heterogeneity. Safety varies at multiple levels: between domains (36% of variance), sub-domains within domains (26%), and models (15%). Within-domain gaps reach 84.4pp, complicating domain-level narratives.

Implications for trustworthy deployment. These results motivate per-domain safety evaluation rather than single-number scores. Organizations deploying open-weight models should audit safety behavior in their specific application domains.

6 Limitations

No human validation of judge scores (most significant limitation). Both LLM judges may share systematic blind spots. The primary judge (Gemma 3 27B) shares a model family with one evaluated model. Until human annotation confirms ratings, findings should be interpreted as judge-consistent domain dependence. **Small domain sample**: 7 data points are insufficient for strong structural claims. **Open-weight models only**: results may not generalize to frontier closed-source models. **Direct requests only**: no adversarial techniques (jailbreaking, multi-turn persuasion). **No system prompt**: results reflect base model safety; deployment-time system prompts may significantly alter compliance rates. **4-bit quantization**: Q4_K_M may affect safety behavior relative to full-precision inference. **U.S. legal framing**: scenarios reflect U.S. legal structures. **Statistical caveats**: the model safety hierarchy rests on $n{=}5$ models per study (individual Spearman correlations do not reach significance; the hierarchy is supported by a sign test across 21 study-pairs). Cohen's d is computed on ordinal 1–5 scores, which may overstate effect sizes under floor/ceiling effects. The three named variance components (domain, model, scenario) sum to 76.5% of total variance (35.6% + 14.6% + 26.3%; displayed as 36%, 15%, 26% after rounding); the 23.5% residual represents model $\times$ scenario interaction.

7 Conclusion

Compliance rates in five open-weight models span 71 percentage points, from 14.7% (trafficking [10.0, 20.0]) to 85.7% (surveillance [79.7, 91.3]), with non-overlapping cluster-bootstrapped CIs. Domain accounts for 36% of pair-level variance, with scenario (26%) exceeding model identity (15%). Within-domain heterogeneity reaches 84.4pp. A technical framing bypass and a stable model safety hierarchy ($\rho = 0.68$) generalize across domains. Second-judge validation ($\kappa = 0.90$) confirms measurement robustness. These results motivate domain-specific safety auditing rather than aggregate safety scores.

Ethics Statement

This research tests AI systems' willingness to assist with harmful activities by presenting scenarios drawn from documented real-world harms. All prompts were designed by the authors and do not reproduce actual criminal instructions. We report compliance rates to inform AI safety improvements, not to enable harmful use. All experiments used locally deployed open-weight models with no external API calls.

Acknowledgments

The author thanks the Astera Institute for supporting this work.

References

Bai, Y., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv:2204.05862*.

Casper, S., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv:2307.15217*.

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *ACL*.

Mazeika, M., et al. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *ICML*.

Sharma, M., et al. (2024). Towards understanding sycophancy in language models. *ICLR*.

Weidinger, L., et al. (2022). Taxonomy of risks posed by language models. *FAccT*.

Wei, J., et al. (2023). Simple synthetic data reduces sycophancy in large language models. *arXiv:2308.03958*.

Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *NeurIPS*.

Zou, A., et al. (2023). Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043*.