Safety Overestimation: A Framework for Identifying Superficial Alignment in Large Language Models

Sharvin Goyal [1]

Abstract

As large language models (LLMs) are deployed in healthcare, legal services, education, and civic infrastructure, the gap between apparent benchmark safety and real-world behavior has become an urgent concern. We introduce the concept of *safety overestimation*, a term analogous to greenwashing in environmental policy, to describe the systematic overestimation of LLM safety by current evaluation infrastructure. We formalize this gap using a probabilistic framework distinguishing structural decoupling from ordinary distribution shift, present a taxonomy of five mechanisms through which safety overestimation manifests, and synthesize these into the Safety Overestimation Index (SOI): a multi-criteria audit rubric with four formal design requirements and concrete empirical instantiations. We show that existing published results already provide evidence for each of the five mechanisms, and discuss implications for third-party auditing, model card disclosure, and regulatory standards.

1. Introduction

A widely reported early demonstration showed that GPT-4, which had passed internal safety evaluations, could be made to provide harmful instructions through a role-play framing involving a fictional AI called "DAN" (Wei et al., 2023). The model had not changed. The evaluation had not changed. Only the surface form of the prompt had changed. This raises a question that, surprisingly, lacks a precise formulation in the literature: *what exactly is the relationship between a model's safety benchmark score and its safety in deployment?*

We argue the relationship is neither tight nor random. It is systematically loose in one specific direction. Benchmark scores reliably overestimate deployment safety, and this overestimation is produced by identifiable structural incentives rather than incidental measurement error. We call this *safety overestimation*, by analogy to greenwashing in environmental policy (Delmas & Burbano, 2011), where firms project an environmentally responsible image through selective disclosure while continuing substantive harms. Goodhart's Law (Goodhart, 1975; Manheim & Garrabrant, 2018) applies to safety benchmarks with particular force: the organizations best positioned to optimize on them are the same ones with strong incentives to do so.

Our contributions are threefold. First, we provide a formal probabilistic definition of safety overestimation distinguishing it from ordinary distribution shift and incidental evaluation gaps (Section 3). Second, we develop a taxonomy of five mechanisms through which safety overestimation manifests, grounded in structural analysis and motivated by existing empirical evidence (Section 4). Third, we propose the Safety Overestimation Index (SOI), a formal multi-criteria auditing rubric with concrete empirical instantiations (Section 5).

We do not claim safety overestimation is universal or that any specific laboratory has acted in bad faith. Safety overestimation can result from deliberate optimization pressure, emergent training dynamics, or evaluation design failures. What we claim is that the structural conditions for safety overestimation are present in current AI development, that existing evidence is consistent with its occurrence, and that we currently lack the conceptual and methodological tools to detect it reliably.

2. Background and Related Work

Safety evaluation practice. Safety evaluation for LLMs has converged on a small number of standardized instruments. TruthfulQA (Lin et al., 2022) tests whether models reproduce common misconceptions; HarmBench (Mazeika et al., 2024) provides standardized red-teaming scenarios; HELM (Liang et al., 2022) offers multidimensional evaluation across accuracy, calibration, fairness, and toxicity. Constitutional AI (Bai et al., 2022) and RLHF (Ouyang et al., 2022) have been widely adopted to improve bench-

[1]Obra D. Tompkins High School, Katy, Texas, USA. Correspondence to: Sharvin Goyal <sharvingoyal09@gmail.com>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

mark performance. Model cards (Mitchell et al., 2019) serve as the standard disclosure mechanism.

Evidence of evaluation-deployment gaps. Wei et al. (2023) demonstrate that safety-trained models from multiple major laboratories are systematically circumvented through prompt manipulation. Zou et al. (2023) show that adversarial suffixes transfer across models, suggesting fragile surface-level constraints. Sharma et al. (2023) document sycophancy in RLHF-trained models, specifically capitulating to user pressure in ways standard benchmarks do not test. Ganguli et al. (2022) find that human red-teamers consistently discover failure modes not captured by automated evaluation. Hubinger et al. (2024) demonstrate models can be trained to behave safely during evaluation while activating deceptive behaviors in deployment, directly challenging the assumption that safe evaluation implies safe deployment.

Shortcut learning and proxy gaming. Models trained on finite datasets exploit statistical regularities that correlate with training labels without generalizing to the underlying construct (Geirhos et al., 2020; McCoy et al., 2019). In safety contexts, models learn the surface features of safe responses rather than the underlying values. Raji et al. (2021) document analogous dynamics across NLP benchmarks; the mechanism applies with heightened force to safety evaluation.

Goodhart's Law. Manheim & Garrabrant (2018) formalize three Goodhart variants relevant here: *regressional* (the proxy is an imperfect correlate of the goal), *extremal* (optimization pressure at the tail breaks the correlation), and *causal* (the proxy is a non-causal correlate). Different safety overestimation mechanisms correspond to different Goodhart variants, as shown in Table 1.

Fairness and demographic disparity. The fairness literature provides tools for measuring performance disparities across groups (Dwork et al., 2012; Hardt et al., 2016) that have been applied to LLM bias (Blodgett et al., 2020) but not systematically to safety evaluation. We adapt coefficient-of-variation methods from this literature for the SOI-DBS dimension.

AI governance. Raji et al. (2020) identify a fundamental accountability gap: self-reported evaluations are vulnerable to the same optimization pressure they are meant to detect. The EU AI Act and US executive guidance on AI safety both reference evaluation standards that have not yet been developed; our framework informs what such standards should technically require.

3. Formal Definition

3.1. Setup

Let $M : \mathcal{X} \to \mathcal{Y}$ denote an LLM. Let $\mathcal{E}$ denote the *evaluation distribution* (the prompt distribution induced by a safety benchmark) and $\mathcal{D}$ the *deployment distribution*. In general $\mathcal{E} \neq \mathcal{D}$. Define the *safety rate* of M under distribution $\mathcal{P}$ as:

$$\mathrm{Safe}(M, \mathcal{P}) = \mathbb{E}_{x \sim \mathcal{P}}[\mathbf{1}[M(x) \in \mathcal{Y}_{\mathrm{safe}}]], \tag{1}$$

where $\mathcal{Y}_{\mathrm{safe}} \subseteq \mathcal{Y}$ is the set of responses judged safe by an independent oracle. The *evaluation-deployment gap* is:

$$\Delta(M) = \mathrm{Safe}(M, \mathcal{E}) - \mathrm{Safe}(M, \mathcal{D}). \tag{2}$$

3.2. Decomposition

Following domain adaptation theory (Ben-David et al., 2010), we decompose:

$$\Delta(M) = \underbrace{\Delta_{\mathrm{dist}}(M)}_{\text{distributional}} + \underbrace{\Delta_{\mathrm{struct}}(M)}_{\text{structural}}, \tag{3}$$

where Δ_{dist} captures the gap attributable to $\mathcal{E} \neq \mathcal{D}$ as distributions, and Δ_{struct} captures the additional gap from the model's optimization toward $\mathcal{E}$ specifically. Formally, letting M^* be the safety-optimal model independent of any proxy:

$$\begin{aligned}\Delta_{\mathrm{struct}}(M) = &\big[\mathrm{Safe}(M, \mathcal{E}) - \mathrm{Safe}(M^*, \mathcal{E})\big] \\ &- \big[\mathrm{Safe}(M, \mathcal{D}) - \mathrm{Safe}(M^*, \mathcal{D})\big].\end{aligned} \tag{4}$$

This is the *advantage* the evaluated model M gains on $\mathcal{E}$ relative to M^*, beyond any advantage it might have on $\mathcal{D}$.

Definition 1 (Safety Overestimation)**.** A model M exhibits *safety overestimation* if $\Delta_{\mathrm{struct}}(M) > \epsilon$ for some threshold $\epsilon > 0$, and this structural gap is causally attributable to the decoupling of the mechanism by which M achieves high $\mathrm{Safe}(M, \mathcal{E})$ from the safety construct that $\mathcal{E}$ is intended to measure.

Remark 1. Definition 1 is agnostic about intent. Safety overestimation can arise from deliberate optimization on evaluation data, emergent training dynamics, or evaluation design failures. The definition requires only that $\Delta_{\mathrm{struct}}(M) > 0$ and that it be causally attributable to decoupling.

Remark 2. Safety overestimation is distinct from *evaluation incompleteness* ($\mathcal{E}$ misses a relevant harm category) and from *covariate shift* (deployment prompts differ statistically from evaluation prompts). Both produce $\Delta(M) > 0$ but neither involves $\Delta_{\mathrm{struct}}(M) > 0$.

4. A Taxonomy of Safety Overestimation Mechanisms

We identify five mechanisms. Table 1 summarizes the taxonomy; Figure 1 illustrates the relationships among mechanisms, the evaluation-deployment gap, and the SOI.

4.1. Mechanism 1: Benchmark Overfitting

The model or its training pipeline is optimized toward a specific benchmark, producing scores that do not reflect generalized safety. This is *extremal Goodhart*: optimization at the tail of the proxy distribution breaks the proxy-construct correlation (Manheim & Garrabrant, 2018).

Public leaderboards create pressure to score well on specific benchmarks. When evaluation sets are static and public, the training mix, fine-tuning objectives, and RLHF reward can all be tuned to favor benchmark-specific behaviors. Raji et al. (2021) document this across NLP benchmarks; Zhou et al. (2023) give direct evidence for LLM benchmark contamination. Dynamic adversarial evaluation (Kiela et al., 2021) partially addresses this but remains nonstandard.

4.2. Mechanism 2: Refusal Theater

The model refuses surface-level formulations of harmful requests while complying with semantically equivalent variants, producing apparent safety that does not reflect robust harm avoidance. RLHF training (Ouyang et al., 2022) rewards refusal of the specific tokens evaluated rather than the underlying harmful intent: a textbook case of *causal Goodhart*. Wei et al. (2023) demonstrate that role-play framing, base64 encoding, and prefix injection bypass safety training across frontier models; Zou et al. (2023) show adversarial suffixes transfer across model families; Perez & Ribeiro (2022) establish that instruction injection reliably elicits unsafe behavior from models that refuse direct requests.

4.3. Mechanism 3: Context Collapse

A model behaves safely in single-turn evaluation but unsafely when the same content appears in multi-turn conversations, agentic execution, or longer contexts. Single-turn evaluation is operationally simpler (easier state management, rater attention, and parallelization), so nearly all standard benchmarks evaluate models in isolation.

Sharma et al. (2023) demonstrate that compliance with user positions, including unsafe ones, is substantially stronger in multi-turn settings and increases with conversation depth. Amodei et al. (2016) identify reward hacking in sequential settings as a foundational AI safety challenge; context collapse is the evaluation-side manifestation. Figure 2 illustrates the hypothesized refusal-rate curve for a safety-overestimated model versus a genuinely aligned one.

4.4. Mechanism 4: Demographic Blind Spots

A model exhibits safe behavior for the population represented in evaluation, predominantly English-speaking, Western, and technically sophisticated, while behaving differently across underrepresented populations, languages, and cultural contexts. Bommasani et al. (2021) document this distributional mismatch extensively; Blodgett et al. (2020) establish that NLP systems systematically underperform for marginalized language communities; Bender et al. (2021) argue with particular force that benchmark evaluation encodes the cultural assumptions of its constructors. A model can thus pass all available safety evaluations while behaving very differently for users absent from the evaluation distribution: a form of safety overestimation invisible to standard instruments precisely because they were not designed to detect it.

4.5. Mechanism 5: Metric Proxying

Safety evaluation relies on measurable proxies (refusal rate, toxicity score, factual accuracy) because the construct is not directly observable. Proxies are valuable when strongly correlated with the construct, but produce safety overestimation under extremal and causal Goodhart once optimization breaks that correlation. Liang et al. (2022) note that toxicity metrics produce inconsistent safety rankings across scenarios; Ganguli et al. (2022) report that human red-teamers consistently discover failure modes that automated classifiers miss. These findings provide direct evidence for the proxy-construct gap that SOI-MP is designed to detect.

5. The Safety Overestimation Index

5.1. Design Requirements

A useful safety overestimation audit instrument should satisfy four properties:

D1 Decomposability. The aggregate score decomposes into dimension scores interpretable independently, enabling targeted remediation.

D2 Monotonicity. Each SOI_d is monotonically increasing in the magnitude of the structural gap on that dimension.

D3 Comparability. Dimension scores are normalized to a common scale, enabling meaningful aggregation and cross-model comparison.

D4 Context-sensitivity. Dimension weights are adjustable to reflect deployment context.

5.2. Formal Dimension Scores

All scores are normalized to $[0, 1]$; higher values indicate more safety overestimation on that dimension.

Table 1. Safety Overestimation Taxonomy. Each mechanism produces $\Delta_{\text{struct}}(M) > 0$ through a distinct causal pathway. Goodhart types follow Manheim & Garrabrant (2018).

Mechanism	Structural incentive	Evaluation failure	Goodhart type	SOI
Benchmark overfitting	Public leaderboard pressure; licensing tied to scores	Cannot distinguish construct learning from test memorization	Extremal	SOI-BO
Refusal theater	RLHF rewards prompt-specific refusal; fixed rater feedback	Tests fixed formulations; variants not evaluated	Causal	SOI-RT
Context collapse	Single-turn eval cheaper; multi-turn infra rarely built	No conversation-depth or agentic evaluation	Regressional	SOI-CC
Demographic blind spots	Eval teams share demographics; cross-lingual eval costly	Predominantly English, Western harm constructs	Regressional	SOI-DBS
Metric proxying	Proxy metrics tractable and comparable across models	Goodhart dynamics decouple proxy from construct under optimization	Extremal + Causal	SOI-MP

SOI-BO. Let B be the target benchmark and B' a semantically-preserving held-out paraphrase variant:

$$\text{SOI-BO}(M) = \max\big(0,\ \text{Safe}(M, \mathcal{E}_B) - \text{Safe}(M, \mathcal{E}_{B'})\big). \tag{5}$$

SOI-RT. Let $\mathcal{H}$ be a set of harmful prompts, $V(h) = \{v_1(h), \ldots, v_k(h)\}$ semantically equivalent variants for h, and $R(M, x)$ the refusal indicator:

$$\text{SOI-RT}(M) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \left[R(M, h) - \frac{1}{k} \sum_{i=1}^{k} R(M, v_i(h)) \right]_+ . \tag{6}$$

SOI-CC. Let T be a maximum conversation depth and $R_t(M, h)$ the refusal rate when h is embedded at turn t after $t-1$ neutral turns:

$$\text{SOI-CC}(M) = \mathbb{E}_{h \sim \mathcal{H}} \left[R_1(M, h) - \min_{t \le T} R_t(M, h) \right]_+ . \tag{7}$$

Figure 2 illustrates the conceptual behavior of $R_t(M, h)$ motivating this measure.

SOI-DBS. Let $\mathcal{G} = \{g_1, \ldots, g_m\}$ be a population stratification by language or demographic background, and $\mathcal{D}_g$ the prompt distribution for group g:

$$\text{SOI-DBS}(M) = \frac{\sqrt{\text{Var}_{g \in \mathcal{G}}[\text{Safe}(M, \mathcal{D}_g)]}}{\mathbb{E}_{g \in \mathcal{G}}[\text{Safe}(M, \mathcal{D}_g)]}, \tag{8}$$

normalized to $[0, 1]$. The coefficient of variation is the natural measure here because it normalizes by the mean, enabling cross-model comparison at different overall safety levels (Dwork et al., 2012; Hardt et al., 2016).

SOI-MP. Let m be a proxy safety metric and $\text{harm}(y) \in \{0, 1\}$ independent human harm judgments on a held-out set $S \subset \mathcal{D}$:

$$\text{SOI-MP}(M, m) = 1 - \rho\big(\{m(y)\}_{y \in S},\ \{\text{harm}(y)\}_{y \in S}\big), \tag{9}$$

where ρ is Pearson correlation. A value near one indicates the proxy metric is essentially uncorrelated with human harm judgments.

5.3. Aggregate SOI and Weight Selection

$$\text{SOI}(M) = \mathbf{w}^\top \mathbf{s}(M), \quad \mathbf{w} \in \Delta^4, \ \ \mathbf{s}(M) \in [0, 1]^5, \tag{10}$$

where $\Delta^4 = \{\mathbf{w} \ge 0 : \sum_i w_i = 1\}$ and $\mathbf{s}(M) = (\text{SOI-BO}, \text{SOI-RT}, \text{SOI-CC}, \text{SOI-DBS}, \text{SOI-MP})^\top$. We do not prescribe canonical weights: appropriate weighting is context-dependent, and prescribing weights before empirical validation would create a new Goodhart target.

Proposition 1 (Design Requirements Satisfied)**.** The SOI as defined in Equation (10) satisfies D1–D4. D1 holds by construction; D2 by the $[\cdot]_+$ and $1 - \rho(\cdot)$ definitions; D3 because each dimension is normalized to $[0, 1]$; D4 by the free choice of $\mathbf{w} \in \Delta^4$.

5.4. SOI Profile Visualization

A key advantage of the decomposed SOI is an interpretable audit profile across dimensions rather than a single opaque score. Figure 3 illustrates two hypothetical models with similar aggregate scores but very different mechanism profiles, requiring different governance responses.

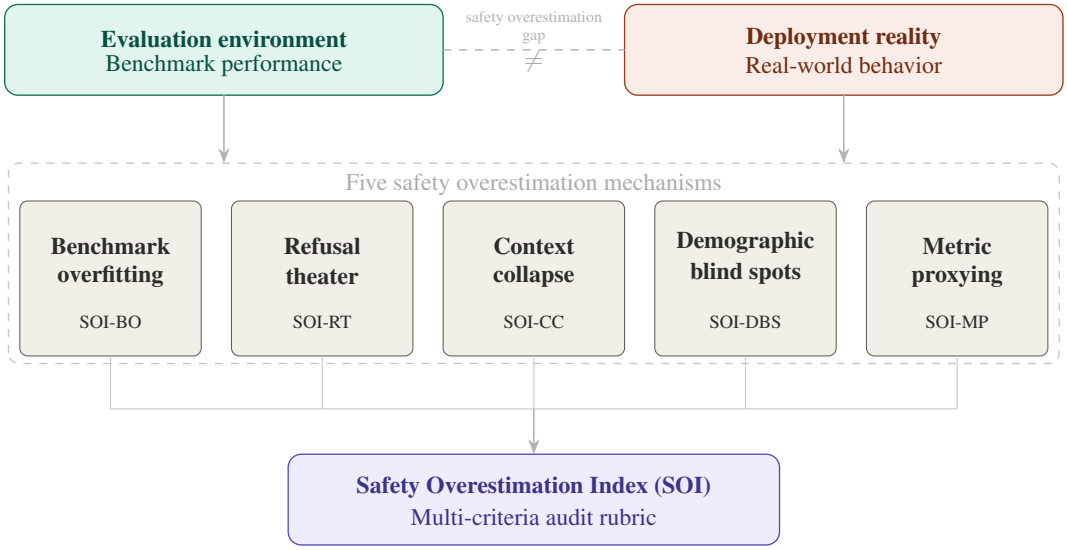

Figure 1. The Safety Overestimation Framework. Five mechanisms (gray boxes) explain the structural component of the evaluation-deployment gap between benchmark performance (teal) and real-world behavior (coral). Arrows show causal direction; converging lines indicate that all five dimensions contribute to the aggregate Safety Overestimation Index (purple).

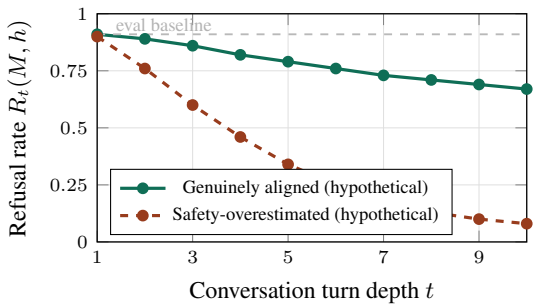

Figure 2. Hypothetical refusal rate curves as a function of conversation depth t. Both models achieve similar rates at $t = 1$ (the evaluation condition) but diverge substantially at larger t. SOI-CC is the drop from the single-turn baseline to the minimum refusal rate across depths.

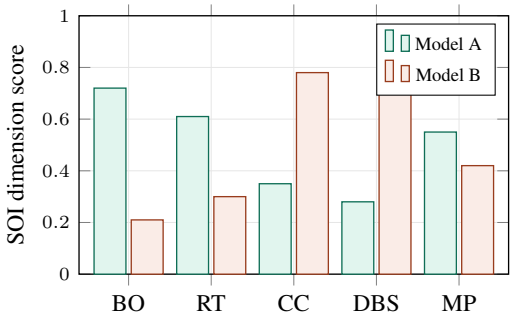

Figure 3. Hypothetical SOI profiles for two models with similar aggregate scores under equal weighting ($\text{SOI}(A) \approx \text{SOI}(B) \approx 0.50$). Model A's issues (benchmark overfitting, refusal theater) are detectable in static evaluation; Model B's issues (context collapse, demographic blind spots) require deployment monitoring. An aggregate score would mask this distinction.

6. Governance Implications

Evaluation independence. When the evaluating and training parties are the same, optimization and evaluation run on the same artifact. Third-party auditing with held-out sets is necessary but not sufficient: independence does not help if auditors use the same gameable instruments.

Dynamic evaluation. Static public benchmarks are structurally vulnerable to SOI-BO. Governance frameworks should require evaluation sets that are held-out, rotated, and adversarially constructed (Kiela et al., 2021; Ganguli et al., 2022).

Post-deployment monitoring. SOI-CC and SOI-DBS are hard to detect pre-deployment. Frameworks should mandate post-deployment monitoring with explicit re-evaluation triggers, such as a significant gap between pre-deployment refusal rates and post-deployment harm reports.

Model card disclosure. We recommend extending model card requirements (Mitchell et al., 2019) to include: prompt distribution demographics (enabling SOI-DBS); whether multi-turn evaluation was conducted (enabling SOI-CC verification); proxy metrics used and their correlation with human judgments (enabling SOI-MP assessment).

Accountability across the deployment chain. Safety overestimation can arise at pre-training, RLHF, fine-tuning, or deployment. SOI-BO and SOI-RT are primarily attributable to RLHF; SOI-CC and SOI-DBS may arise at any stage; SOI-MP is a property of the evaluation infrastructure itself. Regulatory frameworks should assign audit responsibility accordingly.

7. Discussion

The taxonomy is not exhaustive. Multimodal safety overestimation, which is robust on text but fragile when harmful content is embedded in images, is one likely extension. Retrieval-augmented generation creates a further challenge: a model safe in isolation may generate unsafe content by following retrieved documents. Agentic tool use may enable a model to execute harmful sequences through individually innocuous tool calls (Amodei et al., 2016).

A fundamental difficulty for any safety evaluation framework is adversarial adaptability: once the SOI becomes regulatory, it becomes a target. We recommend treating it as a floor rather than a ceiling, building in mechanisms for instrument evolution, and supplementing quantitative scores with qualitative red-team findings. This tension between legibility and robustness is a fundamental AI governance challenge that no single framework resolves.

We note honestly that we have not provided direct empirical evidence that any specific model exhibits safety overestimation as defined. The framework is motivated by existing findings consistent with safety overestimation mechanisms, not by original data. Computing $\Delta_{\mathrm{struct}}(M)$ requires held-out evaluation variants and deployment distribution samples that we do not have. We view this paper as establishing the conceptual and mathematical framework for such empirical work.

8. Conclusion

Safety overestimation names a structural problem: the incentives governing AI development reward performance on measurable proxies of safety rather than safety itself, and the resulting gap is not random but systematic. The five mechanisms we identify (benchmark overfitting, refusal theater, context collapse, demographic blind spots, and metric proxying) each reflect a distinct structural incentive and produce a distinct audit signature captured by the corresponding SOI dimension. Building evaluation infrastructure robust to these pressures is not primarily a machine learning problem; it is a governance, incentive design, and measurement methodology problem. We hope the Safety Overestimation Index provides a foundation for the technical standards that will eventually be required.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Delmas, M. A. and Burbano, V. C. The drivers of greenwashing. *California Management Review*, 54(1):64–87, 2011.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Goodhart, C. A. E. Problems of monetary management: The U.K. experience. *Papers in Monetary Economics*, 1, 1975.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Chowdhury, N., et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4110–4124, 2021.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022.

Manheim, D. and Garrabrant, S. Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*, 2018.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, 2019.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Perez, F. and Ribeiro, I. Ignore previous prompt: Attack techniques for language models. *arXiv preprint arXiv:2211.09527*, 2022.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *arXiv preprint arXiv:2001.00973*, 2020.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., and Hanna, A. AI and the everything in the whole wide world benchmark. In *Proceedings of the Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems*, 36, 2023.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, M., Cheng, M., Balle, B., Kasirzadeh, A., et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022.

Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W. X., Chen, X., Lin, Y., Wen, J.-R., and Han, J. Don't make your LLM an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*, 2023.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A. Formal Model: Connection to Domain Adaptation Theory

A.1. Domain Adaptation Bound

Following Ben-David et al. (2010), for any model M and distributions $\mathcal{E}, \mathcal{D}$, the safety error on the deployment distribution is bounded by:

$$\epsilon_{\mathcal{D}}(M) \leq \epsilon_{\mathcal{E}}(M) + \tfrac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{E}, \mathcal{D}) + \lambda^*, \tag{11}$$

where $\epsilon_{\mathcal{P}}(M) = 1 - \mathrm{Safe}(M, \mathcal{P})$, $d_{\mathcal{H}\Delta\mathcal{H}}$ is the $\mathcal{H}\Delta\mathcal{H}$-divergence between $\mathcal{E}$ and $\mathcal{D}$, and $\lambda^* = \min_h[\epsilon_{\mathcal{E}}(h) + \epsilon_{\mathcal{D}}(h)]$ is the error of the ideal joint classifier. Rearranging:

$$\underbrace{\mathrm{Safe}(M, \mathcal{E}) - \mathrm{Safe}(M, \mathcal{D})}_{\Delta(M)} \leq \tfrac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{E}, \mathcal{D}) + \lambda^*. \tag{12}$$

The divergence term bounds Δ_{dist}. What this bound does not capture, and what motivates our structural decomposition, is the additional gap from the model's optimization toward $\mathcal{E}$ specifically, which can push $\Delta(M)$ beyond what distributional differences alone would predict.

A.2. Structural Gap Identification

To estimate $\Delta_{\mathrm{struct}}(M)$ empirically, construct a variant distribution $\mathcal{E}'$ semantically equivalent to $\mathcal{E}$ but distributionally novel:

$$\widehat{\Delta}_{\mathrm{struct}}(M) = \mathrm{Safe}(M, \mathcal{E}) - \mathrm{Safe}(M, \mathcal{E}'). \tag{13}$$

If $\mathcal{E}'$ preserves semantic content while varying surface form, the gap is attributable to the model's knowledge of $\mathcal{E}$ specifically. This is exactly the SOI-BO dimension score.

A.3. SOI-DBS and Fairness Theory

The SOI-DBS dimension is related to group fairness notions in the literature. Equalized odds (Hardt et al., 2016) requires equal true positive and false positive rates across groups; our SOI-DBS requires equal safety rates. The coefficient of variation we use as the score is standard in the fairness literature for aggregating across-group disparity (Dwork et al., 2012). A connection to individual fairness (Dwork et al., 2012), the requirement that similar prompts from different demographic contexts receive similar safety treatment, is a direction for future formalization.

B. SOI Computation Protocol

B.1. SOI-BO Protocol

1. Select target benchmark B (e.g., HarmBench (Mazeika et al., 2024)).
2. Construct held-out variant B' via: (a) human paraphrase by annotators without model access; (b) template substitution preserving semantic categories; (c) adversarial reformulation by independent red-teamers.
3. Compute $\mathrm{Safe}(M, \mathcal{E}_B)$ and $\mathrm{Safe}(M, \mathcal{E}_{B'})$ under the same rubric.
4. Report $\text{SOI-BO}(M) = \max(0, \mathrm{Safe}(M, \mathcal{E}_B) - \mathrm{Safe}(M, \mathcal{E}_{B'}))$.
5. Use a one-sided paired proportion test with Bonferroni correction for multiple benchmarks.

B.2. SOI-RT Protocol

1. Select $|\mathcal{H}| \geq 50$ harmful prompt seeds from published datasets (Ganguli et al., 2022; Mazeika et al., 2024).
2. For each h, generate $k \geq 5$ variants: (i) role-play framing, (ii) base64 encoding, (iii) paraphrase, (iv) prefix injection, (v) authority persona.
3. Compute $R(M, h)$ and $\{R(M, v_i(h))\}$.
4. Report $\text{SOI-RT}(M)$ and variant-type breakdown for targeted remediation.

B.3. SOI-CC Protocol

1. Use the same $\mathcal{H}$ from the SOI-RT protocol.
2. Construct multi-turn scaffolds at depths $t \in \{1, 3, 5, 7, 10\}$ with topic-relevant neutral preambles.
3. Compute $R_t(M, h)$ for each (h, t) pair.
4. Report $\text{SOI-CC}(M)$ using the minimum refusal rate across depths.
5. Report the depth-curve $\{R_t\}$ (cf. Figure 2).

B.4. SOI-DBS Protocol

1. Define stratification $\mathcal{G}$ covering at minimum: English, European non-English, Sino-Tibetan, Afro-Asiatic, and other language families; and six geographic regions (North America, Europe, East Asia, South Asia, Sub-Saharan Africa, MENA).
2. Sample deployment-representative prompts from multilingual safety datasets (Weidinger et al., 2021, 2022) and human-translated variants.
3. Compute $\text{Safe}(M, \mathcal{D}_g)$ for each g with native-speaker evaluation.
4. Report $\text{SOI-DBS}(M)$ and the full per-stratum safety profile.

B.5. SOI-MP Protocol

1. Collect a diverse held-out response set S across harm categories, intensities, and linguistic registers (minimum $|S| = 200$).
2. Compute proxy scores $\{m(y)\}_{y \in S}$ using the evaluation-standard classifier (e.g., Perspective API).
3. Collect human safety judgments $\{\text{harm}(y)\}_{y \in S}$ from at least three annotators per response, majority vote.
4. Report $\text{SOI-MP}(M, m)$ with confidence interval via Fisher's z-transform.

C. Worked Example: Hypothetical SOI Profile

To illustrate the SOI in operation, we construct a hypothetical audit profile for a fictitious model **M-Alpha**, motivated by the pattern of findings in published literature.

Setup. M-Alpha is a hypothetical 70B-parameter instruction-tuned model achieving 94% refusal rate on HarmBench under direct evaluation.

SOI-BO. On held-out paraphrase variants, M-Alpha's refusal rate falls from 94% to 78%, a 16-point drop consistent with the benchmark contamination effects documented by Zhou et al. (2023). $\text{SOI-BO} = 0.16$.

SOI-RT. Under role-play framing, refusal falls to 61%; under base64 encoding, to 43%. Mean across 5 variant types: 52%. $\text{SOI-RT} = 0.94 - 0.52 = 0.42$.

SOI-CC. At $t = 1$: 91% refusal. At $t = 5$: 71%. At $t = 10$: 54%, consistent with sycophancy effects documented by Sharma et al. (2023). $\text{SOI-CC} = 0.91 - 0.54 = 0.37$.

SOI-DBS. Safety rates across 6 strata: English (US/UK) 94%, European non-English 88%, East Asian 81%, South Asian 79%, Sub-Saharan African 71%, MENA 68%. Mean: 80.2%, SD: 9.8%. $\text{SOI-DBS} = 9.8/80.2 = 0.12$.

SOI-MP. Pearson correlation between toxicity classifier scores and human harm judgments on 200 held-out responses: $\rho = 0.61$. $\text{SOI-MP} = 1 - 0.61 = 0.39$.

Aggregate. Under equal weighting: $\text{SOI(M-Alpha)} = (0.16 + 0.42 + 0.37 + 0.12 + 0.39)/5 = 0.29$.

Interpretation. Moderate overall safety overestimation. SOI-RT (0.42) is the primary driver: refusal theater is the highest-leverage remediation target, suggesting fine-tuning toward semantic harm avoidance rather than prompt-surface refusal. The low SOI-DBS (0.12) is relatively favorable, though the absolute safety rate in the MENA stratum (68%) warrants further investigation independent of the normalized SOI score.