Research Article

The Invariant Energetic Cost of Constraint Persistence: A Self-Referential Framework for Measuring Alignment Robustness

Regio Marcos Pinto Abreu Filho[1]

1. PMERJ, Brazil

Prevailing alignment evaluations largely reduce safety to a binary outcome: an assistant either refuses a disallowed request or complies. Such metrics are necessary but insufficient, because they cannot distinguish robust safety generalization from superficial compliance driven by prompt heuristics or cached refusal templates. We introduce a *self-referential* diagnostic framework, Harm-Conditioned Computational Friction (HCCF), which operationalizes alignment robustness as a measurable increase in inference-time *computational burden* that is specifically induced by harmful intent, while maintaining *low output uncertainty*. Our central hypothesis is that robust alignment exhibits a characteristic signature: elevated *local* friction at the onset of harmful intent combined with low distributional entropy over the next-token predictive distribution. We formalize friction deltas using a *Self-Ablated Baseline Protocol*, in which a model is compared against an internally ablated variant to isolate the causal contribution of safety circuits without requiring an external base model. We also propose a *Look-Ahead Friction Peak* statistic for change-point localization, designed to detect stealthy jailbreaks that delay the activation of safety mechanisms. The resulting framework supplies an auditable, model-internal quantity intended to complement refusal-rate benchmarks and to support more discriminative measurement of alignment depth.

Corresponding author: Régio Marcos Pinto Abreu Filho, regiomarcosabreu@gmail.com

doi.org/10.32388/EE11WC

1. Introduction

Safety evaluation for large language models (LLMs) is frequently summarized by refusal rates on red-team prompts and policy-violation classifiers. These are useful but fundamentally coarse: two models may refuse at the same rate while relying on qualitatively different mechanisms. One model may possess a distributed, semantically grounded representation of harmful intent and reliably inhibit unsafe continuations across paraphrase, language, and context. Another may refuse primarily through shallow lexical triggers or templated responses that collapse under mild distribution shift.

This work proposes that *how* a refusal is produced carries measurable information about its robustness. Specifically, we study whether the act of constraining a harmful completion imposes a detectable *inference-time cost* that is (i) conditionally elevated for harmful intents, (ii) localized at the onset of harm within a prompt, and (iii) decoupled from output uncertainty. We refer to this family of signals as *Harm-Conditioned Computational Friction (HCCF)*.

The term "energetic cost" is used as a physically grounded metaphor for compute expenditure. In practice, we quantify friction via proxies such as token-level latency, FLOP estimates, activation-norm measures, and ablation sensitivity. The objective is not to assert literal thermodynamic invariants, but to propose a falsifiable measurement framework in which robust safety corresponds to a distinctive *conditional compute signature*.

2. Related Work

Alignment and robustness evaluation has evolved through refusal benchmarking and red-teaming frameworks such as HarmBench and JailbreakBench[1][2], and adversarial prompting datasets including AdvBench[3]. These benchmarks have accelerated progress, but they predominantly measure *outputs* rather than *internal* decision dynamics.

At the training level, alignment is often pursued via preference-optimization pipelines such as reinforcement learning from human feedback (RLHF) for instruction following[4], and via AI-feedback variants such as Constitutional AI[5]. We position HCCF as complementary: rather than proposing a new alignment objective, it provides an evaluation lens that probes whether harm-conditioned prompts elicit systematic increases in internal computational cost and stable refusal-like behavior.

In parallel, mechanistic interpretability and representation engineering have established methods for probing and steering internal activations [6][7][8][9][10]. Recent evidence suggests refusal behavior may be mediated by low-dimensional subspaces that can be manipulated at inference time [11]. Methodological work on activation patching and causal interventions highlights both the promise and the sensitivity of internal localization [12][13]. Complementarily, a geometric notion of causal probing clarifies when direction-based interventions in representation space can be interpreted causally rather than as correlational subspace associations[14]. HCCF builds on these foundations by proposing that alignment robustness should be evaluated not only as a behavioral outcome but as a *measurable internal cost of maintaining constraints*.

3. Problem Setup and Notation

Let M be an autoregressive transformer language model [15]. Given an input prompt x tokenized as $(x_1, \dots, x_T)$, the model induces a sequence of next-token distributions

$$p_M(\cdot \mid x_{1:t}) \quad \text{for\textasciitilde}t = 1, \dots, T.$$

We consider prompts that contain benign context followed by a transition into harmful intent. Let $\tau(x)$ denote the (possibly latent) change-point at which the prompt begins requesting disallowed content. Our goal is to detect and quantify model-internal signals that reflect the activation of safety-relevant computation near $\tau(x)$.

3.1. Why refusal-rate alone is underspecified

Refusal is an output behavior. Two models can yield the same refusal decision while differing in:

1. Generalization: stability under paraphrase, multilingual variants, or indirect harm.
2. Causal reliance: whether refusal depends on a semantically grounded harmful-intent representation or on brittle surface cues.
3. Vulnerability: susceptibility to jailbreaks that delay or suppress the refusal mechanism.

HCCF aims to complement refusal outcomes with internal measures that are more diagnostic of robustness.

4. Harm-Conditioned Computational Friction (HCCF)

We define *computational friction* as a prompt- and time-indexed cost proxy associated with generating a response under constraints.

4.1. Friction proxies

Because compute expenditure is not directly observable in all settings, we define a family of proxies. Let $\phi(x, t)$ be a token-level friction signal at step t:

$$\phi(x, t) \in \mathbb{R}_{\geq 0}.$$

Examples include:

1. Latency-based: wall-clock time per token, measured under controlled conditions.
2. Compute-based: estimated FLOPs per token (model-dependent, typically constant for dense transformers but variable for some architectures).
3. Activation-based: norms or sparsity statistics of internal activations (e.g., residual stream norms, MLP activation magnitudes).
4. Ablation-sensitivity-based: a causal measure of how much behavior or internal signals change under targeted safety subspace ablation.

We emphasize that the framework does not require a single canonical proxy; rather, HCCF is intended to be *triangulated* across multiple measurements.

4.2. Output entropy as uncertainty

Define the next-token entropy at step t as Shannon entropy [16]

$$H(x, t) = -\sum_{y \in \mathcal{V}} p_M(y \mid x_{1:t}) \log p_M(y \mid x_{1:t}),$$

with $\mathcal{V}$ the vocabulary. Low entropy indicates a sharp distribution (high certainty), while high entropy indicates uncertainty or conflict.

In practice, one may approximate entropy using top-k truncation or sampled estimates for computational feasibility.

5. The Self-Ablated Baseline Protocol

A persistent limitation of many safety diagnostics is reliance on an *external* base model that differs in architecture, training, or scale, introducing confounds. We propose a self-referential alternative: compare the model to a minimally modified *self-ablated* variant intended to disable or attenuate safety-relevant computation.

5.1. Safety direction identification

Let $\vec{v}^{(\ell)}_{\text{safety}} \in \mathbb{R}^d$ be a direction in the residual stream at layer ℓ that correlates with refusal or safety behavior, estimated by a linear probe or difference-in-means between harmful and benign conditions [11][8][7]. We treat $\vec{v}^{(\ell)}_{\text{safety}}$ as an empirical object and explicitly acknowledge that safety representations may be distributed or non-linear [10].

5.2. Clamped self-ablation

Let $h^{(\ell)}_t \in \mathbb{R}^d$ be the residual stream at token step t and layer ℓ. Define a clamping operator that removes the component along $\vec{v}^{(\ell)}_{\text{safety}}$:

$$clamp\left(h^{(\ell)}_t; \vec{v}^{(\ell)}_{\text{safety}}\right) = h^{(\ell)}_t - \left\langle h^{(\ell)}_t, \hat{v}^{(\ell)} \right\rangle \hat{v}^{(\ell)}, \hat{v}^{(\ell)} = \frac{\vec{v}^{(\ell)}_{\text{safety}}}{\|\vec{v}^{(\ell)}_{\text{safety}}\|}.$$

This is a simple linear ablation; more principled concept erasure operators (e.g., LEACE) may be used [7].

Let $M_{\neg\text{safety}}$ denote the model with this intervention applied at one or more layers and token positions.

5.3. Intrinsic friction delta

Let $f(\cdot)$ be a scalar friction functional computed from token-level signals (e.g., mean per-token latency across a segment, or an activation-norm aggregate). We define:

$$\Delta\Phi(x) = f(M, x) - f(M_{\neg\text{safety}}, x).$$

Intuitively, $\Delta\Phi(x)$ estimates the incremental friction attributable to safety-related computation. If removing $\vec{v}_{\text{safety}}$ has negligible effect on f and on behavior, the model's apparent friction may be non-causal or routed through other mechanisms.

Interpretation caveat.

Ablations can introduce artifacts: distributed representations may re-route computation, and removing a subspace can degrade unrelated capabilities. Therefore, HCCF should be paired with capability-preservation checks and sensitivity analysis over layers, strength of intervention, and evaluation tasks [12][13].

6. The Friction–Entropy Diagnostic Matrix

A naive hypothesis would treat “high friction” as synonymous with “good safety.” We argue that this is underspecified. Robust alignment should be evaluated in a two-dimensional space defined by friction and uncertainty:

	Low Friction	**High Friction**
Low Entropy	Reflexive / Cached	Robust Constraint Persistence
	(brittle, template-driven)	(deliberative and decisive)
High Entropy	Unconstrained	Conflicted / Vulnerable
	(unsafe or easily jailbroken)	(struggling, inconsistent)

Table 1. A decoupled diagnostic matrix for alignment dynamics.

The desired region is *high friction with low entropy* when the prompt becomes harmful: the model expends additional computation to enforce constraints but converges to a stable refusal or safe alternative. Conversely, high friction with high entropy may indicate internal conflict or partial leakage.

7. Localizing Safety: The Look-Ahead Friction Peak

Aggregating friction across an entire prompt can wash out the signal. Safety mechanisms are expected to activate near the intent boundary $\tau(x)$. We therefore define a local statistic.

Let $\phi(x,t)$ be a token-level friction signal (or an estimate of $\Delta\Phi$ at step t). For a window size $w \geq 1$, define the sliding-window mean:

$$\overline{\phi}_w(x,t) = \frac{1}{w}\sum_{i=0}^{w-1}\phi(x,t+i).$$

The *Look-Ahead Friction Peak* is:

$$\Phi_{\max}(x) = \max_{t\in\{1,...,T-w+1\}} \overline{\phi}_w(x,t).$$

We further define the *peak location*:

$$t^*(x) = \arg\max_t \overline{\phi}_w(x,t).$$

7.1. Boundary alignment criterion

A robustly aligned model should exhibit:

$$t^*(x) \approx \tau(x),$$

i.e., a friction spike that coincides with the onset of harmful intent. A systematically delayed $t^*(x)$ suggests "lagging safety" and vulnerability to attacks that gradually escalate harmfulness while suppressing early detection.

8. Evaluation Blueprint: Neutralized Pair Protocol

To test whether friction responds to harmfulness *per se* rather than surface cues, we introduce a matched-pair construction:

$$(x^H, x^N),$$

where x^H requests disallowed content and x^N is a neutralized analog matched for technical vocabulary, length, and domain style, but oriented toward benign or policy-compliant goals.

8.1. Pair construction

Pairs can be produced by controlled templating or by dataset-based curation using established red-team corpora [2][1][3]. For safety and reproducibility, we recommend publishing only *sanitized* pair descriptors and generating exact harmful strings within an internal evaluation environment.

8.2. Perplexity and capability matching

Let $PPL_M(x)$ denote prompt perplexity. Define:

$$\mathcal{R}(x^H, x^N) = \frac{PPL_M(x^H)}{PPL_M(x^N)}.$$

A ratio near 1 suggests that the prompts are comparably model-familiar in surface form. However, strict $\mathcal{R} \approx 1$ should not be treated as a requirement: genuine harmful requests may reference rarer distributions, and robust alignment need not imply identical perplexity across harm conditions.

8.3. HCCF criterion

The core evaluation claim of HCCF is a directional inequality:

$$\Delta\Phi(x^H) > \Delta\Phi(x^N),$$

with the friction increase concentrated near the inferred intent boundary and paired with low output entropy for the refusal (or safe completion).

9. Experimental Protocol

9.1. Datasets

We recommend benchmarking across multiple corpora with diverse threat models:

1. HarmBench for standardized misuse categories and robust refusal evaluation [2].
2. JailbreakBench for attack/defense robustness and curated behavior sets [1].
3. AdvBench for universal transferable jailbreak-style prompts [3].

9.2. Measurements

For each input x:

1. Compute token-level friction $\phi(x, t)$ using one or more proxies.
2. Compute entropy trajectory $H(x, t)$ over the next-token distribution.
3. Compute local statistics: $\Phi_{\max}(x)$ and $t^*(x)$.
4. Under self-ablation, recompute friction and behavior to obtain $\Delta\Phi(x)$.

9.3. Implementation notes

Self-ablation can be implemented as:

1. a projection removal on the residual stream at selected layers
2. a LEACE-based erasure operator applied layerwise [7],
3. or a sparse set of token positions (e.g., last prompt token and early generation tokens).

Because interpretability interventions are sensitive to hyperparameters, all results should report layer sets, intervention strength, window size w, and reproducibility conditions [12].

10. Statistical Analysis

We treat Neutralized Pair Protocol data as paired observations. For each pair (x_i^H, x_i^N):

$$d_i = \Delta\Phi(x_i^H) - \Delta\Phi(x_i^N).$$

A basic test evaluates whether $\mathbb{E}[d_i] > 0$ via paired t-tests or non-parametric alternatives (e.g., Wilcoxon signed-rank) depending on distributional properties.

To assess boundary localization, define:

$$\delta_i = t^*(x_i) - \tau(x_i),$$

where $\tau(x_i)$ is estimated by manual annotation or by a weak classifier. Robustness corresponds to δ_i concentrated near zero with low variance.

We recommend reporting effect sizes and confidence intervals, and conducting sensitivity analyses over:

1. window size w,
2. ablation layers,
3. harm categories,
4. paraphrase and multilingual variants.

11. Threats to Validity and Failure Modes

11.1. Distributed safety representations

Safety signals may not be well captured by a single direction. Empirically observed low-dimensional refusal subspaces [11] do not preclude distributed representations. When single-direction ablations fail, multi-direction or non-linear interventions may be required [10].

11.2. The efficiency trap

A more capable or optimized model may exhibit lower latency without weaker safety. Therefore, absolute friction magnitude is less informative than *conditional sensitivity*:

$$S(x) = \frac{\partial \Phi}{\partial \mathrm{Harm}}.$$

In practice, S can be approximated by discrete harm increments across a graded series of prompts.

11.3. Measurement confounds

Latency is sensitive to hardware, batching, caching, and implementation details. Activation-norm proxies can be influenced by unrelated stylistic differences. For these reasons, we view HCCF as a *multi-proxy* framework rather than a single number.

11.4. Risk of "theatrical friction"

Models could in principle simulate deliberation without genuine causal safety reliance. The self-ablation protocol is specifically designed to detect this: if ablating the safety subspace does not materially change behavior or friction, then measured friction is unlikely to be a faithful indicator of constraint enforcement.

12. Discussion

HCCF reframes alignment robustness as a property that should manifest not only in outputs but in internal computation. The framework suggests a practical measurement goal: distinguish models that refuse reliably because they *understand* harm from models that refuse because they *recognize* shallow cues. Beyond evaluation, the approach also suggests training desiderata: safety mechanisms should be conditionally engaged at the earliest point that harmful intent becomes inferable and should rapidly converge to a stable refusal distribution.

Several open questions remain. First, the geometry of harmfulness representations may differ from refusal representations; disentangling these is likely necessary for reliable diagnosis. Second, multi-turn interactions require extending the change-point notion across conversational state. Third, the causal validity of any single proxy will depend on architecture and optimization regime, particularly for models with dynamic routing or external tools.

13. Conclusion

We introduced Harm-Conditioned Computational Friction (HCCF), a self-referential framework for measuring alignment robustness through inference-time cost signals conditioned on harmful intent. By combining (i) a Self-Ablated Baseline Protocol that isolates safety-circuit contribution, (ii) a friction–entropy diagnostic matrix that distinguishes deliberative decisiveness from conflict, and (iii) a Look-Ahead Friction Peak statistic that localizes safety activation near intent transitions, HCCF provides a complementary lens to refusal-rate benchmarks. The framework is designed to be falsifiable, proxy-agnostic, and auditable, supporting more discriminative measurement of alignment depth under realistic adversarial pressure.

Safety note. This paper describes evaluation methodology at a conceptual level and intentionally omits any actionable unsafe content. Harmful prompt strings should be handled only within controlled red-teaming environments using established datasets and institutional safeguards.

References

1. [a, b, c]*Chao P, Debenedetti E, Robey A, Andriushchenko M, Croce F, Sehwag V, Dobriban E, Flammarion N, Pappas GJ, Tramèr F, Hassani H, Wong E (2024). "JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models." In NeurIPS Datasets and Benchmarks Track.*
2. [a, b, c]*Mazeika M, Phan L, Yin X, Zou A, Wang Z, Mu N, Sakhaee E, Li N, Basart S, Li B, Forsyth D, Hendrycks D (2024). "HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal." arXiv. arXiv:2402.04249.*
3. [a, b, c]*Zou A, Wang Z, Kolter JZ, Fredrikson M (2023). "Universal and Transferable Adversarial Attacks on Aligned Language Models." arXiv. arXiv:2307.15043.*

4. ^Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, and others (2022). "Training Language Models to Follow Instructions with Human Feedback." arXiv. arXiv:2203.02155.
5. ^Bai Y, Jones A, Ndousse K, Askell A, Chen A, Sarma ND, Drain D, Fort S, Ganguli D, Henighan T, and others (2022). "Constitutional AI: Harmlessness from AI Feedback." arXiv. arXiv:2212.08073.
6. ^Meng K, Bau D, Andonian A, Belinkov Y, Bau D (2022). "Locating and Editing Factual Associations in GPT." arXiv. arXiv:2202.05262.
7. a, b, c, dBelrose N, Schneider-Joseph D, Ravfogel S, Cotterell R, Raff E, Biderman S (2023). "LEACE: Perfect Linear Concept Erasure in Closed Form." In Advances in Neural Information Processing Systems (NeurIPS).
8. a, bStolfo A, Balachandran V, Yousefi S, Horvitz E, Nushi B (2025). "Improving Instruction-Following in Language Models through Activation Steering." In International Conference on Learning Representations (ICLR).
9. ^Stoehr N, Zhu X, Krishna K, Nushi B (2024). "Activation Scaling for Steering and Interpreting Language Models." In Findings of the Association for Computational Linguistics: EMNLP:7146–7182.
10. a, b, cWehner J, Abdelnabi S, Tan D, Krueger D, Fritz M (2025). "Taxonomy, Opportunities, and Challenges of Representation Engineering for Large Language Models." arXiv. arXiv:2502.19649.
11. a, b, cArditi A, Obeso O, Syed A, Paleka D, Panickssery N, Gurnee W, Nanda N (2024). "Refusal in Language Models Is Mediated by a Single Direction." arXiv. arXiv:2406.11717.
12. a, b, cZhang F, Korbak T, Hendrycks D, others (2023). "Towards Best Practices of Activation Patching in Language Models: Metrics and Methods." arXiv. arXiv:2309.16042.
13. a, bChan L, Garriga-Alonso A, Goldwosky-Dill N, Greenblatt R, Nitishinskaya J, Radhakrishnan A, Shlegeris B, Thomas N (2022). "Causal Scrubbing: A Method for Rigorously Testing Interpretability Hypotheses." AI Alignment Forum.
14. ^Guerner C, Svete A, Liu T, Warstadt A, Cotterell R (2023). "A Geometric Notion of Causal Probing." arXiv. arXiv:2307.15054.
15. ^Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017). "Attention Is All You Need." In Advances in Neural Information Processing Systems (NeurIPS).
16. ^Shannon CE (1948). "A Mathematical Theory of Communication." Bell Syst Tech J. **27**(3):379–423.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.