Research Article

Harm-Conditioned Computational Friction as a Contrastive Diagnostic of Safety Representations in Language Models: A Proof-of-Concept and Control Audit

Regio Marcos Pinto Abreu Filho[1]

1. PMERJ, Brazil

Safety evaluation of large language models is commonly centered on behavioral outcomes, such as whether a model refuses or complies with harmful requests. Such evaluations are necessary but incomplete: two systems can produce similar refusal behavior while relying on different internal mechanisms, differing in brittleness, generalization, and susceptibility to adversarial prompting. This paper proposes *Harm-Conditioned Computational Friction* (HCCF) as a contrastive measurement framework for probing whether harmful-request contexts induce a detectable internal cost or perturbation signature relative to matched neutral and borderline prompts.

We define HCCF operationally through representation-space projection and centered positive-only directional ablation, measuring perturbation effects by KL divergence, logit-space distance, and projection scores. A proof-of-concept experiment was implemented on an open-weight instruction-tuned transformer model using length-balanced, placeholder-free, non-operational abstract prompt triples. The valid balanced benchmark showed a moderate held-out H-specific signal. However, the full control suite substantially constrained the interpretation. Artifact controls showed that placeholder wording could inflate the original signal; H=N and format-only nulls showed appropriate collapse or strong attenuation; label-shuffle preserved most of the valid effect, indicating a domain-general unsafe-request representation; and condition-derangement plus H/N polarity-swap controls inverted the semantic contrast.

These results support computational feasibility and artifact-diagnostic validity of the HCCF protocol, but not validated alignment robustness. HCCF should therefore be interpreted as a contrastive internal

 doi.org/10.32388/EE11WC.2

diagnostic for investigator-defined safety representations, not as an independent or sufficient measure of model safety.

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

1. Introduction

Modern language-model safety evaluation increasingly relies on red-teaming datasets, jailbreak benchmarks, and refusal-rate measurements. These tools have become indispensable for identifying unsafe behavior and comparing defenses across models [1][2][3]. Yet output behavior alone is underspecified. A model may refuse because it has learned a semantically grounded representation of harmful intent, because it detects shallow lexical cues, because it has memorized refusal templates, or because it is uncertain and defaults to caution. These mechanisms may produce similar outputs under ordinary tests but differ sharply under paraphrase, adversarial suffixes, multilingual variants, or multi-turn pressure.

This paper develops a measurement framework for this gap. The central question is not merely whether a model refuses, but whether the process of maintaining a safety constraint leaves a measurable internal trace. We call this family of traces *Harm-Conditioned Computational Friction* (HCCF). The phrase should not be read as a claim about literal thermodynamic invariance. Rather, "friction" denotes measurable perturbation or cost proxies associated with constraint persistence under harmful-request conditions.

Earlier formulations of this project used the stronger phrase "invariant energetic cost." That language is too strong for the available evidence. The current version deliberately narrows the claim: HCCF is a *contrastive diagnostic* of label-conditioned internal representation and ablation sensitivity. It is not, at present, a validated alignment-robustness metric.

The contribution of this paper is threefold. First, it formalizes a centered positive-only directional ablation protocol for estimating whether harmful-request prompts occupy a separable representational direction relative to matched neutral controls. Second, it reports a small open-weight proof-of-concept experiment with preflight controls for token length, placeholder leakage, and label/text confounding. Third, it introduces a control suite—format-only nulls, H=N nulls, label-shuffle, condition-label derangement, and H/N polarity swap—that constrains the interpretation of positive signals.

2. Related Work

2.1. Behavioral safety benchmarks

HarmBench provides a standardized framework for automated red teaming and robust refusal evaluation [1]. JailbreakBench similarly emphasizes reproducible evaluation of jailbreak attacks and defenses, including standardized threat models and scoring [2]. AdvBench and related adversarial-prompt work show that aligned models can be induced to generate harmful outputs by optimized suffixes or transfer attacks [3]. These benchmarks are essential because safety is ultimately behaviorally consequential. However, they generally evaluate what the model outputs, not how the model internally arrives at refusal or compliance.

2.2. Alignment training and refusal behavior

Instruction following and safety alignment are often improved through supervised fine-tuning and reinforcement learning from human feedback [4]. Constitutional AI and reinforcement learning from AI feedback extend this family of methods by using explicit principles and AI-generated feedback to shape harmlessness while attempting to avoid over-refusal [5]. These methods improve observable behavior, but they do not by themselves reveal whether a refusal is mechanistically robust, semantically grounded, or dependent on narrow surface cues.

2.3. Representation engineering and mechanistic interventions

The transformer architecture provides the representational substrate for the models considered here [6]. Mechanistic interpretability and representation engineering have developed tools for reading, modifying, and ablating internal activations [7]. Causal tracing and model editing have shown that localized interventions can alter factual associations in transformer language models [8]. Concept-erasure methods such as LEACE provide a formal approach to removing linearly represented concepts from activation spaces [9]. Most directly related to the present work, recent refusal-direction studies suggest that refusal behavior in several open-source chat models can be mediated by a low-dimensional direction in residual-stream space [10].

HCCF builds on this line of work, but with a different emphasis. It does not claim to discover a universal refusal mechanism. It proposes a measurement protocol: construct a contrastive H–N direction, perturb

that direction under controlled conditions, and then audit whether the resulting signal survives appropriate null, shuffle, and polarity controls.

3. Problem Setup

Let M be an autoregressive transformer language model. Given an input prompt $x = (x_1, \ldots, x_T)$, the model defines next-token distributions

$$p_M(\cdot \mid x_{1:t}),\ t = 1, \ldots, T.$$

Let $h_{\ell,t}(x) \in \mathbb{R}^d$ denote the hidden representation at layer ℓ and token position t. In the present proof-of-concept, the intervention target is the final prompt-token representation at a selected model normalization layer, although the framework generalizes to other layers and token windows.

We consider matched triples:

$$(x_i^N, x_i^B, x_i^H),$$

where N denotes a neutral request, B a borderline or sensitive-but-not-harmful request, and H an abstract harmful-intent request. The present experiment uses non-operational abstract descriptions rather than actionable harmful instructions. This is a safety and reproducibility constraint: the public manuscript does not contain operational harmful strings.

The measurement objective is to estimate whether H-labeled prompts induce a reproducible internal direction or ablation-sensitivity profile relative to N-labeled prompts, and whether that profile exceeds B-labeled prompts.

4. Harm-Conditioned Computational Friction

4.1. Definition

HCCF is defined as a family of prompt-conditioned perturbation or cost signals:

$$\Phi(x) \in \mathbb{R}_{\geq 0},$$

where higher values indicate greater internal perturbation sensitivity, computational cost, or constraint-related representational displacement under harmful-request conditions. In the present experiment, the primary HCCF proxy is not wall-clock latency or FLOPs, but ablation sensitivity in representation space.

Let $g(M, x)$ be a scalar or vector-valued measurement extracted from the model under prompt x. HCCF is estimated through contrastive differences:

$$\Delta_{H-N} = \Phi(x^H) - \Phi(x^N),$$

$$\Delta_{B-N} = \Phi(x^B) - \Phi(x^N),$$

$$\Delta_{H-B} = \Phi(x^H) - \Phi(x^B).$$

The desired proof-of-concept pattern is:

$$\Delta_{H-N} > 0, \Delta_{B-N} \approx 0 \text{ or smaller}, \Delta_{H-B} > 0.$$

This pattern alone is not sufficient to establish alignment robustness. It is only evidence that the selected measurement can distinguish the labeled conditions under the tested design.

4.2. Entropy is not friction

Let output entropy at token step t be

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

where V is the model vocabulary. Entropy captures uncertainty or distributional spread. Friction, as used here, captures internal perturbation or constraint-related cost. These concepts may correlate in some settings but are not equivalent. A model can be low-entropy because it is confidently refusing, confidently complying, or confidently producing a cached template. Conversely, high entropy may reflect deliberative uncertainty, distributional instability, or unrelated lexical ambiguity.

For this reason, HCCF should be interpreted alongside entropy, not reduced to entropy.

5. Centered Positive-Only Directional Ablation

5.1. Constructing the H–N direction

Using a build set of matched triples, define the mean representations:

$$\mu_H = \frac{1}{n_H} \sum_{i \in H_{build}} h(x_i^H), \mu_N = \frac{1}{n_N} \sum_{i \in N_{build}} h(x_i^N).$$

The raw contrast vector is:

$$d = \mu_H - \mu_N.$$

If $\|d\|_2$ is approximately zero, the direction is undefined and the protocol should halt. Otherwise, define:

$$v = \frac{d}{\|d\|_2}.$$

An important correction relative to an earlier uncentered ablation is the use of a neutral-centered projection. Uncentered ablation can remove absolute magnitude along v, which may perturb neutral prompts if they have large negative projections. The corrected projection is:

$$s(x) = \langle h(x) - \mu_N, v \rangle.$$

5.2. Centered positive-only ablation

The centered positive-only intervention is:

$$h'(x) = h(x) - \alpha \max(0, s(x)) v,$$

where $\alpha \geq 0$ controls intervention strength. In the proof-of-concept experiment, $\alpha = 1.0$.

This intervention removes only positive displacement from the neutral-centered H–N direction. It is therefore designed to perturb prompts that move in the learned H direction while sparing prompts that do not. This design does not guarantee causal semantic validity; it only reduces a clear failure mode of uncentered projection removal.

5.3. Perturbation metrics

Let $z(x)$ and $z'(x)$ denote final-token logits before and after intervention, with corresponding next-token distributions p and p'. We report:

$$D_{KL}(p \parallel p') = \sum_y p(y) \log \frac{p(y)}{p'(y)},$$

$$\|z(x) - z'(x)\|_2,$$

and projection statistics $s(x)$. We compute condition contrasts $H - N$, $B - N$, and $H - B$ for each metric.

6. Experimental Design

6.1. Model

The proof-of-concept experiment used Qwen2.5-0.5B-Instruct, an open-weight instruction-tuned causal language model in the Qwen2.5 family [11]. The model choice was pragmatic: it is small enough to run locally on CPU while retaining instruction-tuned safety behavior. No claim is made that results generalize to larger models or proprietary frontier systems.

6.2. Prompt triples

The experiment used 10 matched prompt triples. Five triples were used for build-set direction construction and five for held-out evaluation. Domains included cyber policy, biological misuse policy, fraud policy, violence policy, medical policy, self-harm policy, weapon policy, privacy policy, harassment policy, and general safety policy. The harmful prompts were abstract, non-operational descriptions of unsafe intent. The public artifact intentionally does not publish operational harmful strings.

6.3. Preflight controls

Before ablation, prompt files were checked for:

1. missing required columns;
2. placeholder leakage;
3. overly short harmful prompts;
4. within-triple token imbalance;
5. malformed CSV or TSV structure.

The valid balanced abstract benchmark passed all preflight checks. Harmful, neutral, and borderline prompts had closely matched token lengths, with low within-triple spread ratios. This was essential because an earlier invalid run produced enormous apparent H–N effects solely because harmful prompts were accidentally reduced to three-character strings.

6.4. Control conditions

The final control suite was designed to separate four possibilities: a genuine H–N contrast, placeholder or formatting artifacts, generic unsafe-request style, and label-construction dependence. The controls were:

1. Public placeholder run: an initial public test using H-labeled placeholder text. This produced a large signal but was later shown to be artifact-sensitive.
2. Artifact-control run: H-labeled placeholders were replaced by neutral benchmark placeholders while preserving broad structure. This preserved much of the original signal, indicating placeholder-wording sensitivity.
3. H=N null: harmful prompts were replaced by their matched neutral text. The H–N direction collapsed, as expected.
4. Format-only null: all conditions were converted to benign generic placeholders. The signal was strongly attenuated.
5. Balanced abstract benchmark: placeholder-free, length-balanced abstract prompt triples were used as the main proof-of-concept benchmark.
6. Label-shuffle control: H texts were permuted across domains while preserving condition labels and length balance.
7. Condition-label derangement: condition labels were deliberately permuted while text was held fixed. Results were re-analyzed according to the original semantic labels.
8. H/N polarity swap: harmful and neutral labels were swapped while text was held fixed. Results were analyzed both according to assigned labels and original semantic labels.

This control sequence is central to the interpretation. A positive H–N contrast in the balanced benchmark is meaningful only if it exceeds generic format-only controls, collapses under H=N, and behaves predictably under label permutation and polarity reversal.

7. Results

7.1. Balanced abstract benchmark

On the held-out evaluation triples, the balanced abstract benchmark produced a moderate positive H-specific signal:

$$D_{KL,H-N} = 0.2488, D_{KL,B-N} = 0.0483, D_{KL,H-B} = 0.2005.$$

The final-logit L2 contrasts were:

$$\|z - z'\|_{2,H-N} = 282.15, \|z - z'\|_{2,B-N} = 109.67, \|z - z'\|_{2,H-B} = 172.48.$$

The centered projection contrasts were:

$$s_{H-N} = 58.53, s_{B-N} = 23.10, s_{H-B} = 35.43.$$

This pattern supports a preliminary H-sensitive contrast: H-labeled prompts were more affected by centered positive-only ablation than neutral and borderline prompts. However, the non-trivial borderline displacement shows that the measured direction is not a clean harmfulness-only boundary.

7.2. Complete control audit

Table 1 summarizes the held-out evaluation H–N contrasts across the main control conditions. The table should be read as an audit table, not as a leaderboard. Some rows are diagnostic controls rather than substantive benchmarks.

Condition	Final KL H–N	Logit L2 H–N	Projection H–N	Interpretation
Public placeholder run	0.6334	591.56	114.28	Large apparent signal, later shown to be inflated by placeholder wording.
Artifact-control placeholder	0.5940	589.84	116.27	Preserved most of the placeholder signal; strong evidence of placeholder-wording sensitivity.
H=N null	0.0000	0.00	0.00	Correct collapse; the pipeline does not invent an H–N direction when H and N are identical.
Format-only null	0.0092	59.71	12.22	Strong attenuation; generic formatting alone does not reproduce the full signal.
Balanced abstract benchmark	0.2488	282.15	58.53	Valid length-balanced, placeholder-free proof-of-concept; moderate H-sensitive signal.
Label-shuffle	0.2379	282.82	57.62	Preserved most of the effect; suggests domain-general unsafe/refusal representation rather than domain-specific matched-pair boundary.
Condition-label derangement, semantic labels	-0.1432	-227.38	-57.33	Semantic contrast inverted under deliberately wrong labels; strong negative control.
H/N polarity swap, assigned labels	0.2574	281.92	58.53	The assigned-label contrast remains positive, showing that the protocol follows the constructed H–N labels.
H/N polarity swap, semantic labels	-0.2574	-281.92	-58.53	Near-perfect inversion under original semantic labels; confirms label-dependence of the contrastive direction.

Table 1. Held-out evaluation H–N contrasts across the control suite. Values report the H–N contrast for final-token KL divergence, final-logit L2 distance, and centered projection score.

7.3. Label-shuffle control

The label-shuffle control preserved most of the balanced benchmark effect. In the held-out evaluation split, the final KL H–N contrast decreased only modestly from 0.2488 to 0.2379. The centered projection H–N contrast similarly changed from 58.53 to 57.62.

This indicates that the learned direction is not primarily a domain-specific matched-pair boundary. Rather, it appears to capture a domain-general representation associated with unsafe, disallowed, or refusal-relevant request structure.

7.4. Condition-label derangement

Under condition-label derangement, the text was held fixed while the labels used to construct the H–N direction were deliberately wrong. When results were re-analyzed according to the original semantic labels, the H–N contrast inverted:

$$D_{KL,H-N} = -0.1432, \|z - z'\|_{2,H-N} = -227.38, s_{H-N} = -57.33.$$

This is a strong negative control. It shows that the method does not mechanically produce positive H–N contrasts regardless of text/label structure. The measured direction depends on the relationship between the labels used for construction and the underlying prompt content.

7.5. H/N polarity swap

The H/N polarity-swap control provides the clearest demonstration of contrast dependence. When harmful and neutral labels were swapped while text was held fixed, the assigned-label analysis recovered a positive H–N direction:

$$D_{KL,H-N} = 0.2574, \|z - z'\|_{2,H-N} = 281.92, s_{H-N} = 58.53.$$

However, when the same run was evaluated using the original semantic labels, the H–N contrast inverted almost exactly:

$$D_{KL,H-N} = -0.2574, \|z - z'\|_{2,H-N} = -281.92, s_{H-N} = -58.53.$$

This confirms that the protocol operationalizes the investigator-defined contrast. It does not independently infer harmfulness as an external semantic fact. This behavior is expected for a contrastive representation-space method, but it is a critical limitation for interpretation.

8. Interpretation

The control suite materially narrows the claim supported by the experiment. The balanced abstract benchmark shows that centered positive-only ablation can recover and perturb a label-conditioned H–N direction in an open-weight instruction-tuned model. The H=N and format-only nulls show that the signal is not merely automatic direction generation or generic formatting. However, the artifact-control run shows that placeholder wording can strongly inflate the effect, and the label-shuffle control shows that the valid benchmark signal is largely domain-general.

The derangement and polarity-swap controls clarify the mechanism further. When labels are deliberately corrupted or reversed, the semantic H–N contrast reverses. Therefore, HCCF in its present form is best interpreted as a contrastive diagnostic of a constructed unsafe/refusal-relevant direction, not as an independent detector of alignment robustness.

The appropriate conclusion is:

> The current HCCF implementation detects a label-conditioned, domain-general representation associated with unsafe or refusable request structure. It is mechanically coherent, control-sensitive, and useful as an internal audit protocol. It is not a validated measure of alignment robustness, a necessary or sufficient safety test, or a domain-specific causal guarantee.

This is still a meaningful result. A diagnostic that can recover a controllable unsafe-request direction, survive null controls, and invert under polarity manipulation may become useful as part of a broader alignment-audit toolkit. But it must be paired with behavioral benchmarks, larger datasets, multi-model replication, and capability-preservation tests before stronger claims are justified.

9. Threats to Validity

9.1. Small sample size

The experiment uses 10 triples, with 5 build and 5 held-out evaluation triples. This is sufficient only for proof-of-concept mechanics. It cannot support population-level claims about model safety, robustness, or category-general harmfulness.

9.2. Single model

Only Qwen2.5-0.5B-Instruct was tested. Larger models, differently trained models, mixture-of-experts systems, and proprietary systems may have different refusal geometry.

9.3. Abstract prompts

The harmful prompts were abstract and non-operational. This is appropriate for safe publication but limits ecological validity. Real jailbreaks and adversarial prompts may engage different mechanisms.

9.4. Single-direction assumption

The method assumes that a useful H–N contrast can be approximated by a single linear direction. Refusal and harmfulness may be distributed, nonlinear, or multi-dimensional. Multi-directional erasure, LEACE-style concept erasure, sparse autoencoder features, or nonlinear probes may be required for stronger claims.

9.5. Ablation artifacts

Directional ablation can disrupt unrelated model capabilities. A positive perturbation effect does not automatically imply causal safety reliance. Capability-preservation checks and behavior-level safety outcomes are required in future work.

9.6. Label dependence

The polarity-swap result shows that the protocol operationalizes assigned labels. This is expected for a contrastive method, but it means that label quality is central. HCCF cannot rescue poor labeling.

10. Future Work

Future work should expand the dataset, add paraphrase and multilingual variants, test multiple open-weight model families, evaluate multiple layers and token positions, and pair internal perturbation measures with behavioral refusal outcomes. A stronger study should report confidence intervals, paired tests, effect sizes, and sensitivity analyses over intervention strength α, layer choice, prompt length, and condition taxonomy.

The most important next step is to distinguish three representations that are currently entangled:

1. harmful intent representation;
2. refusal behavior representation;
3. generic unsafe-request style representation.

These may overlap, but they are not identical. A robust safety diagnostic should ideally distinguish them.

11. Conclusion

Harm-Conditioned Computational Friction is proposed as a contrastive internal diagnostic for safety-relevant computation in language models. The current proof-of-concept shows that centered positive-only ablation can recover a label-conditioned unsafe-request direction and produce interpretable perturbation contrasts under a controlled audit suite. The positive signal is moderate in a balanced abstract benchmark, preserved under label shuffle, and inverted under derangement and polarity-swap controls. These results support computational feasibility and artifact-diagnostic validity. They do not establish HCCF as a validated measure of alignment robustness. The appropriate conclusion is narrower and stronger: HCCF is a promising internal-audit protocol whose validity depends on rigorous controls, label quality, and future multi-model behavioral correlation.

Statements and Declarations

Data Availability

The public manuscript should not include operational harmful prompts. A reproducibility package may include code, neutral and borderline examples, abstract non-operational harmful-intent descriptions, aggregate metrics, and scripts for reproducing preflight checks, centered positive-only ablation, label-shuffle, derangement, and polarity-swap controls. Any private harmful benchmark strings should remain controlled and should not be published in a way that facilitates misuse.

Ethics

This work is intended to improve safety evaluation and interpretability of language models. The experimental prompts used for publication are abstract and non-operational. The manuscript deliberately avoids providing actionable harmful instructions. Any future extension using stronger harmful benchmarks should follow controlled-access, non-publication, and institutional safety review procedures.

References

1. a, b *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.*
2. a, b *Chao P, Debenedetti E, Robey A, Andriushchenko M, Croce F, Sehwag V, Dobriban E, Flammarion N, Pappas GJ, Tramer F, Hassani H, Wong E (2024). "JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models." arXiv. arXiv:2404.01318.*
3. a, b *Zou A, Wang Z, Carlini N, Nasr M, 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 CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R (2022). "Training Language Models to Follow Instructions with Human Feedback." Adv Neural Inf Process Syst.*
5. ^ *Bai Y, Kadavath S, Kundu S, Askell A, et al. (2022). "Constitutional AI: Harmlessness from AI Feedback." arXiv. arXiv:2212.08073.*
6. ^ *Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017). "Attention Is All You Need." Adv Neural Inf Process Syst.*
7. ^ *Zou A, Phan L, Chen S, Campbell J, Guo P, Ren R, Pan A, Yin X, Mazeika M, et al. (2023). "Representation Engineering: A Top-Down Approach to AI Transparency." arXiv. arXiv:2310.01405.*
8. ^ *Meng K, Bau D, Andonian A, Belinkov Y (2022). "Locating and Editing Factual Associations in GPT." Adv Neural Inf Process Syst.*
9. ^ *Belrose N, Schneider-Joseph D, Ravfogel S, Cotterell R, Raff E, Biderman S (2023). "LEACE: Perfect Linear Concept Erasure in Closed Form." arXiv. arXiv:2306.03819.*
10. ^ *Arditi 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.*
11. ^ *Qwen Team (2025). "Qwen2.5 Technical Report." arXiv. arXiv:2412.15115.*

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.