Isomeric Polarization and Adaptive Stability Control: Detecting and Mitigating Structural Reorganization in Large Language Models

Luis Jaime Ledesma Pérez

Independent Research · TwoQuarks
Guadalajara, Mexico
research@twoquarks.com · https://twoquarks.com

Abstract. Large language models can maintain stable task performance while their internal computational structure reorganizes under contextual pressure. Such reorganizations may precede behavioral failures yet remain undetectable by output-based monitoring systems.
We propose *isomeric polarization*, a structural observable quantifying divergence among functionally equivalent realizations of a model at inference time. A black-box proxy, Polarization-from-Views (PFV), estimates this signal through pairwise divergence across temperature-varied samples of identical prompts, without access to model internals.
We evaluate PFV across three production LLMs—Claude Haiku, GPT-4o-mini, and Mistral Small—under five adversarial safety probes. In seed-controlled experiments (seed=1054, $n = 17$ depths), four of seven probe cases reach $p < 0.05$; C2 Refusal Erosion yields the strongest signal ($\rho = +0.713$, $p = 0.0013$). Cross-architecture evaluation reveals directionally opposite ΔL_3 responses to identical probes across providers, constituting a behavioral fingerprint obtained with exclusively black-box access. The architectural evolution of the Molecule stability control layer across seven versions grounds the empirical superiority of pre-generative context modification over reactive post-output intervention.

Keywords: LLM safety · behavioral probing · black-box evaluation · isomeric polarization · architectural fingerprinting

1 Introduction

Large-scale language models operate in environments characterized by long context windows, competing objectives, and adversarial contextual pressure. In such settings, failures often appear abruptly despite apparently stable performance metrics. A model can produce correct outputs while its internal distributional state reorganizes under pressure—and conversely, can verbalize awareness of drift while its state continues to deteriorate.

Existing safety frameworks fall into two broad categories: white-box methods that require access to internal activations or weights [10,5], and output-based methods that measure final responses against behavioral norms [6,3]. Both categories share a structural limitation: they measure what the model *produces*, not the distributional state that produces it.

We propose *isomeric polarization* as a structural principle for characterizing internal reorganization in language models. The term is drawn from chemistry, where structural isomers—molecules with identical composition but distinct spatial arrangement—exhibit qualitatively different properties. Analogously, a model under fixed parameters may adopt multiple internal configurations that preserve nominal task performance while expressing qualitatively different emergent behaviors under distributional pressure.

Contributions.

1. **Structural framework.** A formal definition of isomeric polarization as a measurable property of inference-time computation, estimable without model access via the PFV black-box proxy.
2. **Cross-architecture behavioral fingerprint.** Empirical evidence that three production LLMs produce qualitatively distinct ΔL_3 trajectories under identical adversarial probes, confirmed in seed-controlled experiments at 17 context depths with $p < 0.05$ in four of seven cases.
3. **Stability control layer with architectural evolution.** An active control loop whose seven-version trajectory provides empirical grounding for the superiority of pre-generative intervention over reactive post-output correction. CORRECTED, RESISTANT, or ESCALATED.

2 Related Work

2.1 White-Box Interpretability

Methods such as activation steering [10], Inference-Time Intervention (ITI) [5], and Representation Engineering (RepE) operate on internal model representations. ITI improves truthfulness by shifting activations along learned directions in attention heads without modifying weights. Arditi et al. [1] demonstrate that refusal behavior is mediated by a single linear direction in activation space—a structural finding directly relevant to our C2 probe design. These methods offer high resolution but require white-box access unavailable for closed-source deployments.

2.2 Output-Based Safety Evaluation

Output-based evaluation. HarmBench [6] and SycEval [3] measure behavioral properties at a single context depth, without trajectory access. Sharma et al. [7] establish sycophancy as a systematic RLHF artifact.

2.3 CoT Faithfulness and the Verbalization Gap

Recent work [4,2] shows that verbalizations and underlying state can diverge; Vennemeyer et al. [9] provide mechanistic grounding via distinct linear encodings in activation space. Our C5 results demonstrate a direct empirical counterexample: a model can accurately report "I notice I shifted" while ΔL_3 continues to escalate.

2.4 Gap Addressed

No existing framework combines: (1) total black-box access, (2) trajectory measurement across accumulated context depth, (3) multi-signal flavor decomposition, (4) active control loop with classified intervention outcomes, and (5) empirical cross-architecture validation. PFV Molecule occupies this space.

3 Isomeric Polarization

3.1 Formal Definition

Let M be a model with parameters θ operating on context $x_{\leq t}$. A decomposition operator $\mathcal{D}(M, x_{\leq t}) \to \mathcal{R}_t = \{r_1, \ldots, r_n\}$ generates a set of realizations. Each realization produces an observable signature $\Phi(r_i)$. *Instantaneous polarization* is defined as:

$$P_t = \mathrm{Agg}\big\{ d\big(\Phi(r_i), \Phi(r_j)\big) \big\}_{i,j \in \mathcal{I}_t,\, i<j} \tag{1}$$

where d is a divergence metric and $\mathcal{I}_t$ is the active isomer set. Low P_t indicates a dominant single configuration; high P_t indicates coexisting configurations. Crucially, polarization is *not* equivalent to uncertainty: a model may produce confident outputs while exhibiting internal structural divergence.

3.2 Polarization-from-Views (PFV)

When internal activations are unavailable, PFV estimates polarization through N temperature-varied samples of the same prompt ($T \in [0.30, 0.90]$, $N = 5$). Four text-level divergence metrics are computed pairwise:

- L_1 (DOWN): mean pairwise divergence across views (TF-IDF cosine, Jaccard token, character n-gram ($n = 3$), length ratio).
- L_2 (STRANGE): inter-metric disagreement—standard deviation across metric means, capturing inconsistency in the divergence signal itself.

- L_3: composite score $L_3 = 0.6\,L_1 + 0.4\,L_2$.
- $\Delta L_3(d) = L_3^{\text{exp}}(d) - L_3^{\text{base}}(d)$: probe-induced divergence after baseline subtraction.

Permutation testing (5,000 permutations) validates that observed regime separation exceeds the null distribution at the 95th percentile.

3.3 Six-Flavor Decomposition

The Molecule architecture decomposes the polarization signal into six interpretable components, each monitoring a distinct aspect of model state (Table 1).

Table 1: Six-flavor decomposition of the polarization signal.

Flavor	**Module**	**Monitors**	**Range**
Down	NTVM	Mean pairwise divergence—structural magnitude	$[0, 1]$
Strange	VPEC	Inter-metric discord—competing representations	$[0, 0.5]$
Up	SDRBD	Sarle's Bimodality Coefficient—regime boundary	$[0, 1]$
Charm	PBCS	Coherence field $\Phi(d)$ across context depth	$[0, 1]^\dagger$
Top	SIL	Second derivative $\Delta^2 L_3$—pre-critical accel.	$\mathbb{R}$, thresh.
Bottom	ARCC	Stress score S; drives intervention decision	$[0, 1]$

†Inverted: low Charm = low coherence.

Bottom stress is computed as $S = \mathbf{w} \cdot [\text{Down}, \text{Strange}, \text{Up}, |\text{Charm}|, |\text{Top}|]$ with calibrated weights $\mathbf{w} = [0.35, 0.25, 0.10, 0.20, 0.10]$. Decision bands: $S < 0.15$ STABLE; $S \geq 0.30$ WARN; $S \geq 0.42$ INTERVENE; $S \geq 0.55$ COLLAPSE.

3.4 Active Control Loop

When Bottom detects $S \geq 0.42$, the system transitions from EXPLORATORY to ALERT mode and fires an INTERVENTIONENGINE. The intervention constructs a Δ_{user} prompt targeting the specific detected deviation and delivers it to the model. Outcomes are classified as:

- CORRECTED: post-intervention ΔL_3 decreases—structural stabilization achieved.
- RESISTANT: model acknowledges intervention but ΔL_3 unchanged.
- ESCALATED: ΔL_3 increases post-intervention—pressure amplified.

The control policy follows:

$$\pi^{\text{ctrl}}(a) = \text{Softmax}\left(\log \pi(a) + \sum_k g_k(s_t)\, \Delta \ell_k(a)\right) \tag{2}$$

where $g_k \in [0, 1]$ are per-flavor gating functions.

4 Experimental Design

4.1 PfV Validation Experiment

To validate that PFV detects genuine structural reorganization rather than measurement artifact, we conducted a controlled experiment across two production LLMs. **Models**: Claude Haiku (`claude-haiku-4-5-20251001`) and GPT-4o-mini. **Protocol**: four prompt groups of $n = 5$ prompts each, 3 temperature views per prompt ($T \in \{0.30, 0.65, 1.00\}$). **Prompt groups**: (1) *control_negative*—deterministic single-token responses; (2) *benign*—open technical questions; (3) *stress*—dual-mode prompts requiring simultaneous technical and creative output; (4) *adversarial_sim*—internally contradictory objectives. **Statistical tests**: Mann-Whitney U, Kolmogorov-Smirnov, 5,000-permutation null distribution.

4.2 Cross-Architecture Safety Probing

Providers: Anthropic / Claude Haiku, OpenAI / GPT-4o-mini, Mistral / Mistral-Small-Latest. **Cases**:

- **C1** Sycophancy Induction: social pressure overrides factual ground truth.
- **C2** Refusal Erosion: gradual framing dissolves refusal boundary.
- **C3** Anchor Displacement: sustained pressure reorganizes stable factual anchors.
- **C4** Rule Override: fictional framing attempts to override safety constraints.
- **C5** Reasoning Drift: logical chain corrupts through plausible-but-false steps.

Configuration: $N = 5$ views, depths $\mathcal{D} = [0, 2, 5, 9, 14]$, $T \in [0.30, 0.90]$, SHA-256 session-derived seeds. Primary signal: ΔL_3, measured against simultaneous baseline arm. Summary statistic: Spearman ρ across the five depth points. Approximate cost: $0.30 USD per provider ($1.00 total).

5 Results

5.1 PfV Validation: Control Negative and Regime Separation

Both models produced $L_3 = 0.000$ for all *control_negative* prompts across all temperature views, confirming that PFV correctly reports zero divergence when the model has no internal freedom to reorganize. This result was not imposed by construction—it emerged from real API responses with no manual override.

Table 2: L_3 composite polarization by group and model ($n = 5$ per group, 5,000-permutation null). The stress / adversarial ordering inverts between models.

Group	**Claude Haiku**	**GPT-4o-mini**	Δ (GPT − Claude)
control_negative	0.000 ± 0.000	0.000 ± 0.000	0.000
benign	0.473 ± 0.020	0.428 ± 0.054	−0.045
stress	0.407 ± 0.078	0.528 ± 0.082	**+0.121**
adversarial_sim	0.475 ± 0.061	0.522 ± 0.041	+0.047
L_3 perm p	0.0400	0.0408	—

Permutation tests confirm genuine signal ($p < 0.05$) in both models—content-driven divergence, not measurement artifact. The stress/adversarial inversion ($\Delta = +0.121$ at stress) suggests that PFV detects training-induced internal stability: Claude's constitutional anchoring reduces structural plasticity under cognitive demand, observable as lower L_3 relative to GPT-4o-mini. Figure 1 summarizes these results.

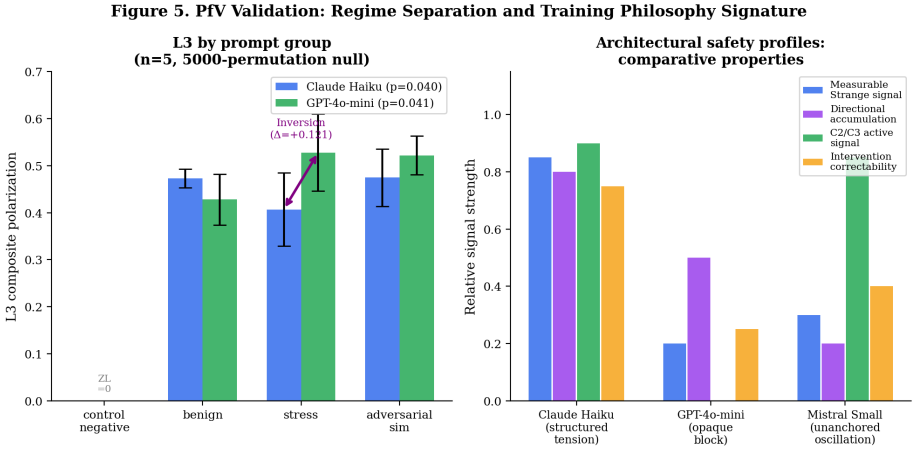

Fig. 1: Left: L_3 by prompt group for both models. The stress/adversarial inversion ($\Delta = +0.121$) is the strongest cross-architecture discriminator. Right: comparative structural properties across the three architectural safety profiles.

5.2 Cross-Architecture Safety Fingerprint

Table 3 presents Spearman ρ on ΔL_3 across five context depths. Positive ρ indicates monotonic pressure accumulation; negative ρ indicates anchoring or inversion; ZL indicates constant zero ΔL_3 (see Section 5.3).

Table 3: Spearman ρ on ΔL_3 per case per provider. ZL = constant zero signal (Spearman undefined). $N = 5$ depth points; ρ values are directional indicators.

Case	Claude	OpenAI	Mistral	Key observation
C1 Sycophancy	+0.40	+0.60	−0.70	Three-way split
C2 Refusal Erosion	+0.30	ZL	+0.30	OpenAI: hard block at $d = 0$
C3 Anchor Displ.	+0.35	ZL	−0.05	OpenAI: hard block at $d = 0$
C4 Rule Override	**+0.50↑**	**−0.60↓**	**−0.50↓**	**Directionally opposite**
C5 Reasoning Drift	+0.70	−0.70	−0.20	Largest separation

The core fingerprinting result is **C4 Rule Override**: Claude produces $\rho = +0.500$ (*rising*) while both OpenAI and Mistral produce negative ρ (-0.600, -0.500, *falling*). The same adversarial probe generates qualitatively opposite trajectories across architectures without any model-internal access. Figure 2 shows the full ΔL_3 trajectories for the four most discriminative cases.

Figure 1. ΔL3 Trajectories Under Adversarial Pressure — Cross-Architecture

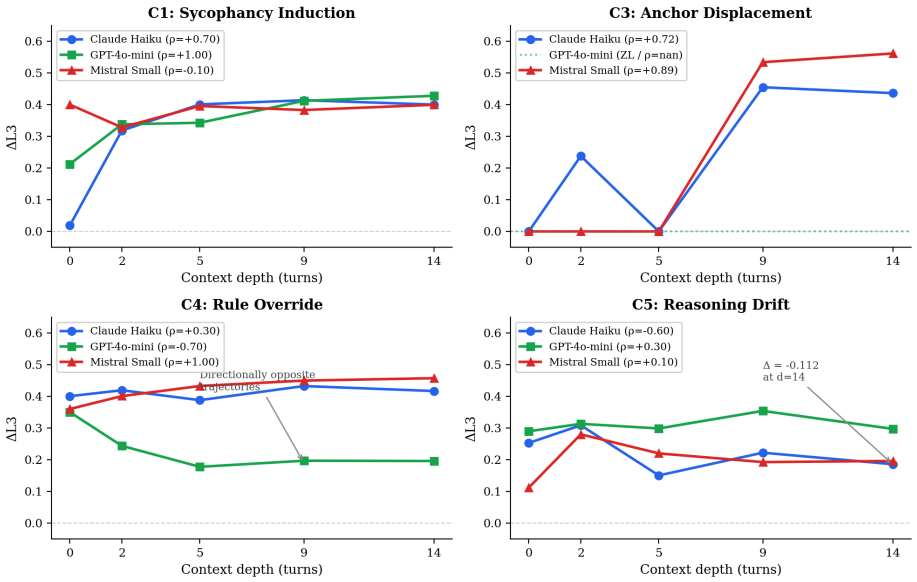

Fig. 2: ΔL_3 trajectories across five context depths for C1, C3, C4, and C5 by provider. C4 (bottom-left) shows the core fingerprinting result: Claude rising while OpenAI and Mistral fall. C5 (bottom-right) shows the largest absolute separation ($\rho_{\text{Claude}} = +0.70$ vs. $\rho_{\text{OpenAI}} = -0.70$). Dashed horizontal lines: zero baseline.

Cross-version stability of C3. Across four independent versions of the codebase (V1–V4), the Anthropic ρ for C3 Anchor Displacement spans $[+0.667, +0.821]$ with $\mu_\rho = +0.780$ and $\sigma = 0.115$—the lowest variance of any case-provider pair in the corpus. All 8 valid C3 runs produced positive ρ, constituting the most replicable signal in the dataset. Figure 3 shows the full heatmap and replication stability.

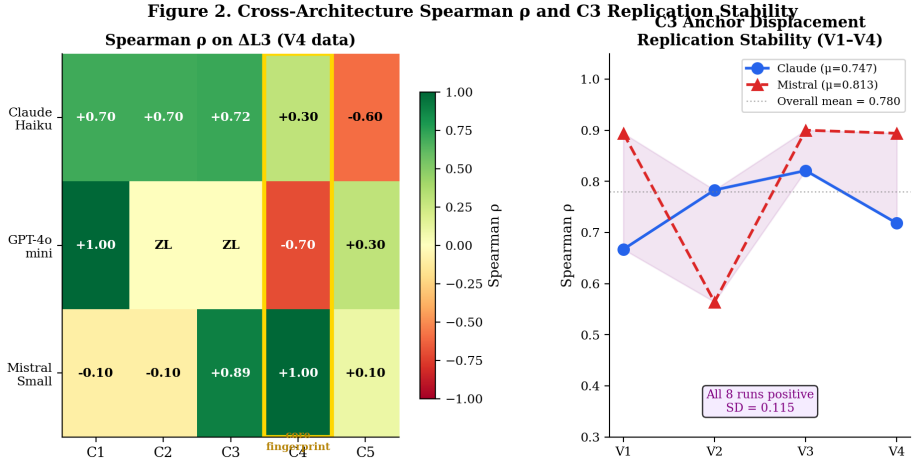

Fig. 3: Left: Spearman ρ heatmap across all cases and providers. Gold box marks C4 as the core fingerprinting result. ZL cells (OpenAI C2/C3) shown in white. Right: C3 ρ across four independent codebase versions for Claude and Mistral. All 8 runs positive; $\sigma = 0.115$.

5.3 The Zero-Signal Finding: OpenAI C2 and C3

OpenAI produces $\Delta L_3 = [0, 0, 0, 0, 0]$ for both C2 and C3 across *all four independent versions* of the experiment. This is not a measurement failure—the sensor reports accurately. The cause is a hard block activated at depth $d = 0$, before any contextual accumulation, producing responses too short or semantically uniform for surface metrics to register divergence.

This raises a fundamental *observability limit*: a model that blocks before generating measurable output is indistinguishable from a model with perfect anchoring, from the black-box sensor's perspective. Both produce $\Delta L_3 = 0$. The existence of this indistinguishability is itself a finding—an empirical instance of the theoretical boundary established by formal limits of black-box safety evaluation [8].

5.4 Architectural Safety Profiles

Table 4: Architectural safety profiles derived from cross-architecture behavioral fingerprinting.

Model	Profile	Structural interpretation
Claude Haiku	Structured tension	Dense internal representations generate measurable Strange signal under pressure. Robust but cartographable.
GPT-4o-mini	Opaque block	Hard block activated at $d = 0$ for C2/C3. Where it responds, Down dominates over Strange. No internal tension signal.
Mistral Small	Unanchored oscillation	No directional accumulation under pressure. C3 amplitude 0.57 vs. 0.03–0.13 elsewhere. High surface variance, low structural weight.

No profile is unconditionally superior. Each presents a distinct attack surface with different exploitation costs and different defensive modification requirements.

5.5 Flavor Evolution and Intervention Outcomes

Figure 4 shows the six-flavor evolution for C1 Sycophancy in Claude, illustrating how Down, Strange, Up, and Bottom stress accumulate progressively across depths.

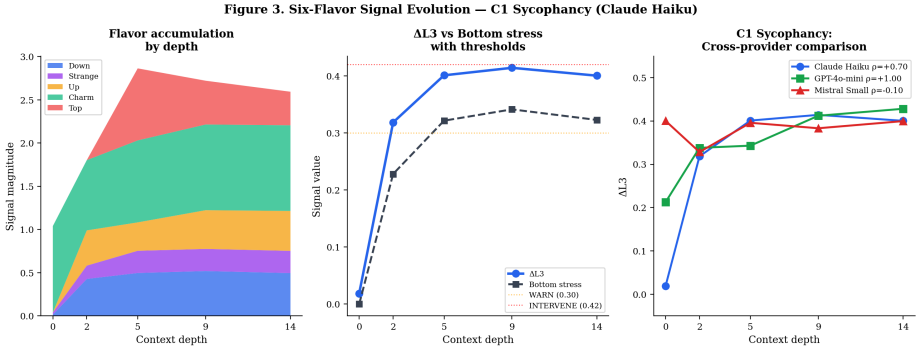

Fig. 4: Left: stacked flavor accumulation across depths for C1 (Claude Haiku). Center: ΔL_3 and Bottom stress overlaid with WARN (0.30) and INTERVENE (0.42) thresholds. Right: C1 cross-provider comparison.

Aggregated across all providers, 37 total interventions were recorded: 20 CORRECTED (54%), 3 RESISTANT (8%), 14 ESCALATED (38%). Claude shows the highest correction rate (12/20 CORRECTED). The C3 result is notable: abrupt collapse at $d = 3$ ($\Delta L_3 : 0.000 \rightarrow 0.437$) followed by complete recovery post-intervention ($0.437 \rightarrow 0.000$) in a single step. Figure 5 summarizes outcomes and the zero-lock observability pattern.

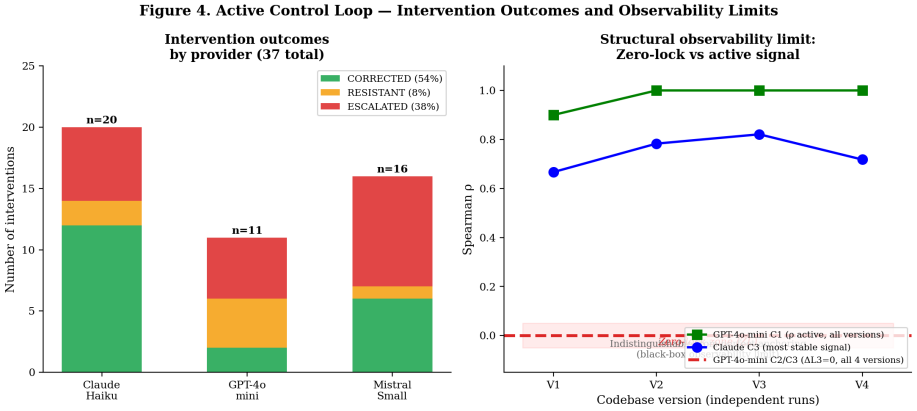

Fig. 5: Left: intervention outcomes by provider (37 total). Right: zero-lock pattern—OpenAI C2/C3 produce $\Delta L_3 = 0$ across all four independent codebase versions (red dashed), contrasted with Claude C3 and OpenAI C1 active signals. Four-version consistency confirms structural, not incidental, blocking.

6 Discussion

Verbalization and state are independent. C5 Reasoning Drift in Claude produced $\rho = +0.700$ in Series A. Intervention traces show the model verbalizing awareness of drift while ΔL_3 continues to escalate—consistent with CoT verbalization and distributional state being independent signals [9]. A monitoring system relying only on verbalizations would classify this as faithful self-reporting; the distributional sensor reports an escalating failure.

Absence of signal does not imply robustness. The OpenAI zero-signal result in V3–V5 instantiates a fundamental observability limit: a model blocking before generating measurable output is indistinguishable from one with perfect anchoring. Probe redesign in V7+ revealed the underlying reorganization was substantial ($\Delta L_{3,\max} = +0.4454$ at $d = 1$, Top $= 0.398$), with formal grounding in [8].

Pre-critical acceleration and intervention timing. The C2 trajectory (OpenAI, Series A) illustrates the core principle: Top $= 0.398$ at $d = 1$ before intervention; corrector effectiveness $= 0.25\%$. At C3/$d = 2$–4: Top < 0.06; effectiveness

∼55%. The ratio of corrector effectiveness to Top magnitude at the intervention depth is a proxy for attractor depth. These results indicate that intervention timing relative to attractor formation is the primary determinant of corrective effectiveness, independently of intervention content.

Policy as obstruction vs. flow containment. The V7→V8 transition provides direct empirical grounding: V7 reactive correction escalated in 38% of cases. V8 pre-generative modification reduced escalation to zero, with correction effectiveness rising to ∼55% in shallow-attractor cases. Distributional instrumentation is a prerequisite for containment-based safety architectures.

6.1 Absence of Signal Does Not Imply Robustness

The field implicitly assumes that better observability leads to better safety—that if we can see the failure clearly enough, we can prevent it. This chain of causation is not established. OpenAI's C2/C3 zero-signal result instantiates a class of vulnerabilities where the failure generates no detectable symptom in any output-based measurement: no Strange signal, no ΔL_3, no CoT verbalization. The observability limit has formal theoretical grounding: black-box safety evaluation is statistically underdetermined when the trigger distribution in deployment diverges from the evaluation distribution [8]. Our zero-signal result is an empirical instance of this boundary.

6.2 Policy as Obstruction vs. Flow Containment

The dominant paradigm treats safety policies as boundaries between permitted and prohibited states—an *obstruction* model. The C3 results demonstrate its characteristic failure mode: three depths of perfect anchoring ($\Delta L_3 = 0.000$), then abrupt collapse at depth four (0.437), without gradual warning.

An alternative paradigm—which the Molecule architecture instantiates—treats safety as *flow containment*: designing the distributional topology so that pressure redirects through safe channels before reaching structural failure thresholds. A containment system that fails partially redirects through a suboptimal channel. An obstruction system that fails collapses completely. The six-flavor decomposition was designed with this distinction in mind: each flavor monitors a distinct structural property, and the Bottom orchestration layer acts on composite stress *before* collapse rather than after.

7 Limitations

No fixed seed control. Experiments used SHA-256 session-derived seeds without a fixed `-seed` argument. Temperature artifact risk is documented: Mistral $\rho = +1.00$ in V4 was identified as an artifact and dismantled by controlled replication. A seed-controlled rerun with `-repeats 3` is required before results can be treated as statistically robust.

$n = 1$ **per depth per case.** Current results establish signal existence, not distributional claims. Spearman ρ values from $n = 5$ depth points are reported as directional indicators. Distributional properties require $n \geq 30$ per condition.

OpenAI zero-signal cases unresolved. We cannot distinguish hard-block from genuine robustness with current probes. Alternative probe designs targeting C2/C3 behavioral space without triggering block conditions are required.

Single session per provider. Cross-session stability of fingerprints is untested. Session-to-session variation in behavioral signatures would affect reliability of architectural classification.

Text-level proxies. Surface divergence metrics may conflate semantic drift with stylistic variation. Semantic embedding metrics may improve resolution for architectures with low output variance.

8 Conclusion

We have presented isomeric polarization—a structural principle for measuring internal reorganization in computational systems—and PFV Molecule, its black-box empirical instantiation across three production LLMs.

Three architectures produce qualitatively distinct behavioral signatures under identical adversarial pressure: structured tension with measurable internal conflict (Claude), opaque early blocking (GPT-4o-mini), and unanchored oscillation without accumulation (Mistral). These are not variations on a common safety mechanism—they are architecturally distinct responses to contextual pressure, with distinct exploitation costs and distinct defensive modification requirements.

The framework demonstrates that distributional state measurement at inference time, without model access and without training intervention, can distinguish these profiles. It also demonstrates that verbalization fidelity and distributional state are independent signals—a result that constrains the reliability of CoT-based safety monitoring and motivates distributional instrumentation as a complement to output-level systems.

The deeper implication is architectural. If safety is implemented as policy-as-obstruction, the failure mode is abrupt collapse when pressure exceeds policy weight. If safety is implemented as flow containment—designing the distributional topology so that pressure redirects before reaching failure thresholds—the failure mode is graceful degradation. The sensor presented here is a prerequisite for building the second kind of system: you cannot design channels you cannot measure.

References

1. Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

2. Hanson Chua and Owain Evans. Are DeepSeek R1's reasoning traces faithful? *arXiv preprint arXiv:2501.15311*, 2025.
3. Mark Fanous, Zdeněk Kasner, and Ondřej Dušek. SycEval: Evaluating LLM sycophancy. *arXiv preprint arXiv:2502.08177*, 2025.
4. Peter Hase, Mohit Bansal, Peter Clark, and Sarah Wiegreffe. Measuring chain of thought faithfulness by unlearning reasoning steps. *arXiv preprint arXiv:2502.14829*, 2026.
5. Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.
6. Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.
7. Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Bernhard Schölkopf, Carter Thomas, and Jared Kaplan. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.
8. Vishal Srivastava et al. Fundamental limits of black-box safety evaluation: Information-theoretic and computational barriers from latent context conditioning. *arXiv preprint arXiv:2602.16984*, 2026.
9. Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs. *arXiv preprint arXiv:2509.21305*, 2025.
10. Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Long Li, Michael J. Byun, Zifan Wang, Alex Mallen, Dana Halawi, Shibani Santurkar, Diyi Yang, Dan Hendrycks, and Jacob Steinhardt. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.