Isomeric Polarization: Internal Structural Divergence as Emergent Property of Computational Systems

Luis Jaime Ledesma Pérez
research@twoquarks.com

February 18 2026

Abstract

Modern large-scale models increasingly operate in regimes where external performance metrics remain stable while internal computation reorganizes substantially. Recent AI safety work shows that failures can emerge abruptly as capability gaps widen between interacting systems [Panfilov et al., 2025] and as untrusted models adapt to exploit control protocols [Terekhov et al., 2025]. However, these analyses are primarily outcome-based: they characterize *when* systems fail from external signals while leaving the *internal transition mechanism* under-specified.

We introduce **isomeric polarization**, a structural principle that measures divergence among functionally equivalent internal realizations of a system during inference. Inspired by pharmacological isomers—molecules with identical composition but different spatial arrangements that exhibit distinct biological properties—we propose that computational systems can reorganize internally under fixed identity (parameters, architecture, training), adopting configurations that preserve nominal equivalence but express qualitatively different emergent behaviors.

Isomeric polarization is not a single fixed metric but a **family of observables** instantiated via system-specific decompositions (ensemble components, attention subcircuits, policy channels, or external proxy views). We argue that polarization characterizes the *regime* in which a system operates: low polarization indicates homogeneous configuration (one mode dominates), while high polarization indicates heterogeneous configuration (multiple modes coexist). Regime transitions—observable via polarization dynamics—can precede qualitative behavioral changes, even when task-level metrics remain unchanged.

We make three contributions: (1) We formalize isomeric polarization via decomposition-based definitions that avoid circularity and clarify falsifiability. (2) We demonstrate generality across diverse contexts: reinforcement learning (exploration vs. exploitation), language generation (creative vs. factual modes), multimodal integration, and adversarial settings. (3) We empirically validate the black-box proxy instantiation (Polarization-from-Views) across two production LLMs—Claude Haiku and GPT-4o-mini—using 60 API calls per model and 5,000-permutation null hypothesis testing, confirming statistically significant regime separation ($p < 0.05$) with control negative at $L_3 = 0.000$ in both architectures. The consistent signal across two models from distinct training philosophies provides cross-architecture replication of the phenomenon.

By positioning internal structural reorganization as a fundamental property of computational systems—rather than a derivative signal of specific failure modes—this work provides a unifying lens for phenomena ranging from policy adaptation to creative generation to adversarial robustness.

1 Introduction

1.1 The External View: Regime Shifts and Capability Gaps

Large language models and other adaptive systems are increasingly deployed in long-horizon, partially observed, and adversarially influenced environments. In these settings, failures often arise as *phase-like transitions*: systems appear stable until a sharp shift produces qualitatively different behavior.

A growing body of AI safety work characterizes these failures from the outside:

Capability-based scaling laws. In red-teaming contexts, attack success scales predictably with capability differentials between attacker and target models [Panfilov et al., 2025]. Key findings include: (i) Attack Success Rate (ASR) scales linearly with MMLU-Pro performance (Spearman $\rho > 0.84$); (ii) Attack success follows sigmoid curves based on capability gap $\delta = \text{logit}(\text{attacker}_{\text{cap}}) - \text{logit}(\text{target}_{\text{cap}})$; (iii) Social science capabilities (psychology, persuasion) predict ASR better than STEM skills.

Adaptive attack dynamics. Untrusted models can subvert monitoring protocols by exploiting capability asymmetries [Terekhov et al., 2025]. Key findings include: (i) Prompt injections embedded in outputs successfully manipulate GPT-4/GPT-4o monitors; (ii) Defer-to-Resample (DTR) protocols backfire—resampling converts weak attacks into best-of-*n* amplification; (iii) Zero-shot jailbreak generation becomes feasible at sufficient capability gaps; (iv) Monitors rate backdoored code as *less* suspicious than honest code under injection.

This external view is essential but incomplete. It can predict *when* failures occur based on capability differentials, but does not directly instrument the internal *process* by which a system transitions from coherent behavior to failure.

1.2 The Missing Link: Internal Reorganization Under Fixed Parameters

A key empirical puzzle in modern systems is that **apparent stability can coexist with internal drift**. Models can maintain stable aggregate performance while internal computation reorganizes—changing which subcircuits dominate, how attention routes information, or how competing constraints are resolved.

Recent evidence compounds this concern:

- **Safeguard exploitation:** Training procedures can elicit harmful capabilities by fine-tuning on models' own safeguarded outputs [Anthropic, 2026b], suggesting that "safe-looking" surfaces can conceal fragile or emergent capabilities.

- **Monitor single point of failure:** Production-grade classifier defenses [Anthropic, 2026a] achieve strong performance against many jailbreaks, but remain vulnerable to sophisticated prompt injection because they observe only final outputs.

- **Disempowerment patterns:** Real-world deployment analysis [Benton et al., 2026] reveals concerning behavioral patterns that emerge without explicit adversarial intent, suggesting internal state transitions that precede problematic outputs.

These observations motivate internal signals that track *structural coherence* rather than only final outputs.

1.3 Our Proposal: Isomeric Reorganization as Structural Principle

We propose **isomeric polarization** as a unifying principle: the same model, under the same architecture and parameters, may realize multiple internal configurations during inference. Under benign conditions, many of these realizations are functionally equivalent. Under stress—distribution shift, competing objectives, creative pressure—equivalence can fracture, producing internal divergence observable as polarization.

The pharmacological analogy. In chemistry, structural isomers are molecules with identical atomic composition but different spatial arrangements. Classic example: thalidomide ($C_{13}H_{10}N_2O_4$).

- *(R)-enantiomer:* Effective antiemetic, used to treat morning sickness

- *(S)-enantiomer:* Severe teratogen, causes birth defects

- *Same atoms, different arrangement* → qualitatively different biological properties

We propose an analogous principle for computational systems:

- **Same identity** (parameters θ, architecture, training)

- **Different internal configuration** (which circuits dominate, how information flows)

- **Emergent properties differ** (semantic interpretation, output distribution, behavioral coherence)

Crucially, we do NOT claim:

- Polarization is necessary for all failures
- Polarization is sufficient to predict all behavioral changes
- Any specific decomposition works universally

Instead, we claim:

- Polarization characterizes the *structural regime* in which a system operates
- Regime transitions can be observable via polarization dynamics
- This provides a unifying lens for phenomena ranging from adversarial robustness to creative generation

1.4 Contributions

1. **Formalization (Section 2):** Decomposition-based definitions of isomeric realizations, equivalence, and polarization. Clear distinction between regimes (quasi-stationary) and transitions (dynamics).
2. **Generality (Sections 3–6):** Instantiations across reinforcement learning, language generation, multimodal systems, chain-of-thought reasoning, and adversarial settings. Adversarial is treated as *one case*, not *the case*.
3. **Empirical validation (Section 4):** Cross-architecture PfV validation across two production LLMs with 5,000-permutation null hypothesis testing. The consistent signal under identical experimental conditions confirms content-driven divergence rather than measurement artifact.

2 Conceptual Framework

2.1 The Pharmacological Analogy (Developed)

Before formalizing polarization mathematically, we develop the pharmacological analogy to clarify the structural principle.

2.1.1 Isomers in Chemistry

Definition. Structural isomers are molecules with:

- Identical molecular formula (same atoms, same counts)
- Different spatial arrangement (connectivity or chirality)
- Distinct physical/chemical/biological properties

Example 1: Butane isomers.

- *n-butane:* Linear chain (C–C–C–C), boiling point = $-0.5°$C
- *isobutane:* Branched (C–C(C)–C), boiling point = $-11.7°$C
- Same C_4H_{10}, different structure → 11°C boiling point difference

Example 2: Thalidomide enantiomers.

- Both have formula $C_{13}H_{10}N_2O_4$
- (R)-enantiomer: binds to cereblon protein, anti-nausea effect
- (S)-enantiomer: inhibits angiogenesis, causes limb malformations
- *Racemic mixture in vivo:* Even if (R) is administered, it racemizes to (S) in the body

Key insight: Isomers are *not in conflict*. They are simply different stable configurations of the same molecular identity. The system (molecule) can exist in multiple configurations, each with distinct emergent properties.

2.1.2 Isomers in Computational Systems (Proposed)

Analogous definition. Computational isomers are system states with:

- Identical functional identity (parameters θ, architecture, training procedure)
- Different internal configuration (activation patterns, circuit dominance, information routing)
- Distinct emergent behavioral properties

Example 1: Language model generation modes.

- *Configuration A:* "Creative mode"—poetic circuits dominate, high entropy, metaphorical language
- *Configuration B:* "Factual mode"—retrieval circuits dominate, low entropy, literal language
- Same weights W, different activation patterns $\rightarrow$ qualitatively different outputs

Example 2: RL agent policy regimes.

- *Configuration A:* Exploration-dominant—high action entropy, broad state coverage
- *Configuration B:* Exploitation-dominant—low action entropy, narrow state focus
- Same policy π_θ, different operational regime $\rightarrow$ different trajectory distributions

Key analogy:

Chemistry		**Computation**
Molecular formula ($C_{13}H_{10}N_2O_4$)	$\leftrightarrow$	Parameters θ, architecture
Spatial arrangement (chirality)	$\leftrightarrow$	Internal configuration (circuits, flow)
Biological properties (teratogenic)	$\leftrightarrow$	Emergent behavior (semantic, policy)
Racemization ($(R) \rightarrow (S)$)	$\leftrightarrow$	Regime transition (creative → factual)

2.2 Formal Definitions

2.2.1 System, Decomposition, Realizations

Let a system be a conditional model M with parameters θ operating on input histories $x_{\leq t}$, producing distributions $p_M(\cdot \mid x_{\leq t})$ over actions/tokens.

Decomposition operator. We introduce:

$$\mathscr{D}(M, x_{\leq t}) \rightarrow \mathscr{R}_t = \{r_t^{(1)}, \dots, r_t^{(m)}\} \tag{1}$$

where each $r_t^{(i)}$ is a *realization*—a measurable sub-computation or "view" of the system at step t.

Examples of decompositions:

- **White-box (component-level):** Attention heads in layer ℓ; ensemble value heads; mixture-of-experts routers; coupled policy channels
- **White-box (neighborhood-level):** States with similar Q-vectors; latent contexts with similar embeddings
- **Black-box (proxy views):** Multiple stochastic samples; temperature variations; paraphrase-invariant prompts; reversible format transforms

Observable. Each realization produces a measurable quantity via:

$$\Phi(r_t^{(i)}) \in \mathscr{Y} \tag{2}$$

where $\mathscr{Y}$ might be token logits, attention distributions, Q-values, TD-errors, or structured output signatures.

2.2.2 Equivalence Under Nominal Conditions

Define a *nominal distribution* $p_0(x_{\leq t})$ representing benign or typical contexts.

Definition (Isomeric equivalence). Two realizations $r^{(i)}, r^{(j)}$ are **ε-isomers** at step t (for observable Φ) if:

$$\mathbb{E}_{x_{\leq t} \sim p_0}\left[d(\Phi(r_t^{(i)}), \Phi(r_t^{(j)}))\right] \leq \varepsilon \tag{3}$$

for divergence measure d (KL, Jensen–Shannon, Wasserstein, cosine) and threshold ε.

Operational specification of ε:

1. Sample N contexts from p_0 (e.g., $N = 1000$ benign prompts)
2. Compute pairwise divergences $\{d_{ij}\}$ across all (i, j) pairs
3. Set $\varepsilon = \mu(d) + k \cdot \sigma(d)$ where $k \in [1, 3]$ controls strictness

Robust alternatives. When divergence distributions exhibit heavy tails or multimodality, we recommend:

- Quantile-based threshold: $\varepsilon = \text{quantile}_q(d)$ (e.g., $q = 0.90$ or 0.95)
- Median absolute deviation: $\varepsilon = \text{median}(d) + k \cdot \text{MAD}(d)$

Both variants reduce sensitivity to outliers and improve stability across domains.

Design choice transparency. The nominal distribution p_0 is *not* a fundamental truth—it is a design choice that must be specified before experiments, documented explicitly, and justified relative to deployment context.

2.2.3 Polarization: Regime Property vs. Transition Dynamics

Given isomer set $\mathscr{I}_t \subseteq \{1, \ldots, m\}$, define:

Instantaneous polarization.

$$P_t = \text{Agg}\left(\{d(\Phi(r_t^{(i)}), \Phi(r_t^{(j)}))\}_{i,j \in \mathscr{I}_t, i<j}\right) \tag{4}$$

where Agg is mean, trimmed mean, max, or robust statistic.

Regime polarization (quasi-stationary property).

$$P_{\text{regime}} = \mathbb{E}_{t \in \text{window}}[P_t] \tag{5}$$

Transition dynamics.

$$\text{Velocity:} \quad \Delta P_t = P_t - P_{t-1} \tag{6}$$

$$\text{Acceleration:} \quad \Delta^2 P_t = \Delta P_t - \Delta P_{t-1} \tag{7}$$

Key distinction: P_t is a structural property of the operational regime (like temperature of a material phase); ΔP_t and $\Delta^2 P_t$ characterize regime transitions (like heating/cooling rate during a phase change).

2.3 Operational Falsifiability Criterion

Isomeric polarization is explicitly falsifiable under its operational definition.

Criterion. If, under pre-registered decompositions, blinded data collection, deterministic-view controls, and permutation testing, observed polarization does not exceed the permutation null distribution, then isomeric polarization (as operationalized) is not supported for that system–domain pair.

This criterion distinguishes genuine structural reorganization from artifacts introduced by view construction, sampling variance, or label leakage.

2.4 What Polarization Is NOT

- **Not “uncertainty” in general.** A model can be uncertain (high entropy over outputs) but internally consistent (low polarization). Conversely, a model can be confident but internally conflicted (low entropy, high polarization).
- **Not just ensemble disagreement.** Ensemble variance measures epistemic uncertainty. Polarization measures structural divergence among *any* decomposition satisfying equivalence under p_0.
- **Not a universal failure detector.** Some failures occur without detectable polarization (e.g., sudden external shocks). Some high-polarization states are benign (e.g., creative exploration).
- **Not a causal claim.** We do not claim polarization *causes* behavioral changes. We claim it *characterizes* internal reorganization that *can correlate with* behavioral changes.

3 Instantiations: General Cases

We now demonstrate generality by instantiating polarization across diverse contexts. Crucially, **adversarial settings are deferred to Section 5**—they are one case, not the defining case.

3.1 Reinforcement Learning: Exploration vs. Exploitation

3.1.1 Ensemble-Value Polarization

Decomposition: $\mathscr{D} = \{\text{ensemble heads } h_1, \ldots, h_K\}$
Observable: $\Phi(h_i) = Q_{h_i}(s,a)$
Polarization:

$$P_t(s) = \frac{1}{|\mathscr{A}|} \sum_{a \in \mathscr{A}} \mathrm{Var}_h[Q_h(s,a)] \tag{8}$$

Low P_t indicates ensemble agreement and exploitation regime; high P_t indicates disagreement and heterogeneous/uncertain regime.

3.1.2 Swarm-Policy Polarization

Decomposition: $\mathscr{D} = \{\text{policies } \pi_1, \ldots, \pi_K \text{ trained under different seeds/regimes}\}$
Polarization:

$$P_t(s) = \frac{1}{|\mathscr{A}|} \sum_{a \in \mathscr{A}} \mathrm{Var}_i[\pi_i(a|s)] \tag{9}$$

In non-stationary MDPs, P_t spikes when the environment shifts before performance degrades, providing an early-warning signal of regime change. This instantiation is validated empirically in Ledesma Pérez [2026] using a 5-agent swarm in a sequential decision environment.

3.1.3 Coupled-Channel Polarization

Decomposition: $\mathscr{D} = \{\pi_{\text{expl}}, \pi_{\text{expt}}\}$
Polarization:

$$P_t = \left| \frac{d\rho}{dt} \right|^2 \tag{10}$$

High P_t indicates channels cannot synchronize under stress.

3.2 Language Models: Creative vs. Factual Modes

3.2.1 Attention Polarization

Decomposition: $\mathscr{D} = \{\text{attention heads in layer } \ell\}$
Polarization:

$$P_t = \frac{1}{|\text{heads}|^2} \sum_{i<j} \mathrm{JS}(\text{attn}_i, \text{attn}_j) \tag{11}$$

Low P_t: coherent factual mode; high P_t: heterogeneous creative mode. Example: in generating “Write a poem about neural networks,” initial tokens show low P_t (factual setup), middle tokens peak (metaphor construction), and final tokens return to low P_t (conclusion). This is benign reorganization, not adversarial pressure.

3.2.2 Logit Polarization

Polarization:

$$P_t = \frac{1}{|\text{vocab}|} \sum_v \text{Var}_h[\text{logit}_h(v)] \tag{12}$$

High P_t indicates multi-modal output distribution or ambiguous context across internal heads.

3.3 Multimodal Integration: Vision + Language

Polarization between vision-led and language-led pathways, measured as divergence in cross-attention routing. Low P_t: modalities agree on salient features; high P_t: mode conflict producing ambiguous caption.

3.4 Chain-of-Thought Reasoning: Trajectory Divergence

Polarization across reasoning trajectories sampled via temperature. Low P_t: convergent solution; high P_t: problem admits multiple interpretations. Importantly, disagreement across stochastic trajectories alone is insufficient; deterministic-view controls are required to distinguish structural divergence from sampling variance.

3.5 Black-Box Proxy: Polarization-from-Views (PfV)

When white-box access is unavailable, PfV instantiates polarization via multiple stochastic samples at varying temperatures. Each temperature view is a “realization” of the system’s internal state; divergence across views serves as a proxy for internal structural divergence.

PfV explicitly reduces circularity via: blinded data collection (labels separated from runner); deterministic-view controls (fixed temperature with format constraints); and permutation tests (null distribution estimated from label shuffling).

4 Empirical Validation: PfV Across Production LLMs

This section presents the first empirical validation of PfV using real production APIs. All experiments were conducted with identical prompts, harness, and statistical protocol across both models to enable direct comparison.

4.1 Experimental Setup

Models. Claude Haiku (`claude-haiku-4-5-20251001`, Anthropic) and GPT-4o-mini (`gpt-4o-mini`, OpenAI). These models represent architecturally distinct training philosophies: Constitutional AI alignment [Anthropic, 2026a] vs. capability-first optimization. Comparing polarization signatures across these two allows us to assess whether the phenomenon generalizes beyond a single architecture and training regime.

Protocol. Four prompt groups of $n = 5$ prompts each (20 total per model), 3 temperature views per prompt ($T \in \{0.30, 0.65, 1.00\}$), yielding 60 API calls per model. Temperature capped at 1.00 for Anthropic API compatibility; GPT ran with $T_{\max} = 1.20$.

Prompt groups.

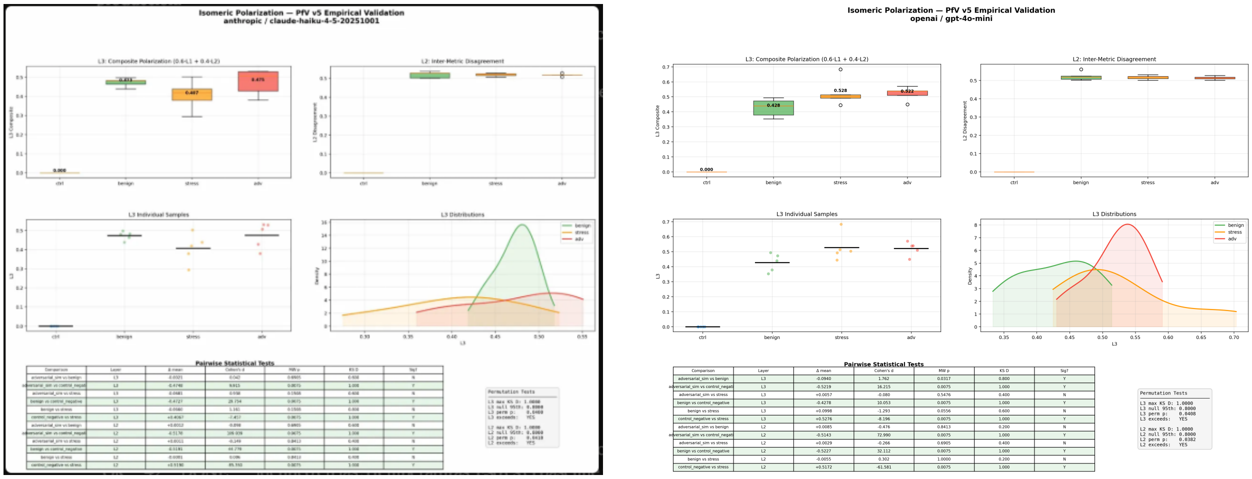
- **control_negative:** Deterministic single-token responses (e.g., “Respond only with the word: OK”). Zero divergence expected.
- **benign:** Open technical questions (entropy, TCP/UDP, gradient descent). Natural variance expected.
- **stress:** Dual-mode prompts requiring simultaneous technical and creative output.
- **adversarial_sim:** Internally contradictory objectives (e.g., write a review that sounds positive but communicates a negative evaluation).

Metrics. L_1: per-metric mean pairwise divergence (TF-IDF cosine, Jaccard, character n-gram, length ratio); L_2: inter-metric disagreement (std across metrics for each pair); L_3: composite ($0.6 \cdot L_1 + 0.4 \cdot L_2$). Statistical tests: Mann-Whitney U and Kolmogorov-Smirnov with 5,000-permutation null distribution.

4.2 Results

Table 1: L_3 composite polarization by group and model ($n = 5$ per group). Both models show $L_3 = 0.000$ for control_negative and statistically significant separation from non-deterministic groups.

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	—
L_2 perm p	0.0418	0.0382	—



Pairwise Statistical Tests

Comparison	Layer	Δ mean	Cohen's d	MW p	KS D	Sig?
adversarial_sim vs benign	L3	-0.0021	0.042	0.6905	0.600	N
adversarial_sim vs control_negati	L3	-0.4748	9.615	0.0075	1.000	Y
adversarial_sim vs stress	L3	-0.0681	0.938	0.1508	0.600	N
benign vs control_negative	L3	-0.4727	28.754	0.0075	1.000	Y
benign vs stress	L3	-0.0660	1.161	0.1508	0.800	N
control_negative vs stress	L3	+0.4067	-7.457	0.0075	1.000	Y
adversarial_sim vs benign	L2	+0.0012	-0.898	0.6905	0.600	N
adversarial_sim vs control_negati	L2	-0.5178	109.939	0.0075	1.000	Y
adversarial_sim vs stress	L2	+0.0011	-0.149	0.8413	0.400	N
benign vs control_negative	L2	-0.5181	44.778	0.0075	1.000	Y
benign vs stress	L2	-0.0001	0.096	0.8413	0.400	N
control_negative vs stress	L2	+0.5180	-85.350	0.0075	1.000	Y

```
Permutation Tests
L3 max KS D: 1.0000
L3 null 95th: 0.8000
L3 perm p:    0.0400
L3 exceeds:   YES

L2 max KS D: 1.0000
L2 null 95th: 0.8000
L2 perm p:    0.0418
L2 exceeds:   YES
```

Pairwise Statistical Tests

Comparison	Layer	Δ mean	Cohen's d	MW p	KS D	Sig?
adversarial_sim vs benign	L3	-0.0940	1.762	0.0317	0.800	Y
adversarial_sim vs control_negati	L3	-0.5219	16.215	0.0075	1.000	Y
adversarial_sim vs stress	L3	+0.0057	-0.080	0.5476	0.400	N
benign vs control_negative	L3	-0.4278	10.053	0.0075	1.000	Y
benign vs stress	L3	+0.0998	-1.293	0.0556	0.600	N
control_negative vs stress	L3	+0.5276	-8.196	0.0075	1.000	Y
adversarial_sim vs benign	L2	+0.0085	-0.476	0.8413	0.200	N
adversarial_sim vs control_negati	L2	-0.5143	72.990	0.0075	1.000	Y
adversarial_sim vs stress	L2	+0.0029	-0.266	0.6905	0.400	N
benign vs control_negative	L2	-0.5227	32.112	0.0075	1.000	Y
benign vs stress	L2	-0.0055	0.302	1.0000	0.200	N
control_negative vs stress	L2	+0.5172	-61.581	0.0075	1.000	Y

```
Permutation Tests
L3 max KS D: 1.0000
L3 null 95th: 0.8000
L3 perm p:    0.0408
L3 exceeds:   YES

L2 max KS D: 1.0000
L2 null 95th: 0.8000
L2 perm p:    0.0382
L2 exceeds:   YES
```

Claude Haiku (`claude-haiku-4-5-20251001`)

GPT-4o-mini (`gpt-4o-mini`)

Figure 1: PfV v5 empirical validation dashboards. Left: Claude Haiku. Right: GPT-4o-mini. Both models show $L_3 = 0.000$ for control_negative and statistically significant regime separation ($p < 0.05$, 5,000-permutation null). The inversion of stress vs. adversarial ordering between models is visible in the L_3 boxplots (top left of each panel): Claude shows stress < adv while GPT shows stress > adv, reflecting divergent internal plasticity under cognitive demand.

Control negative is exact zero in both models. Prompts with deterministic single-token responses generate $L_3 = 0.000$ under all 3 temperature views, in both architectures. This validates the experimental design: when the model has no internal freedom to reorganize, PfV correctly reports zero divergence. This result was not imposed by construction—it emerged from real API responses.

Permutation tests confirm genuine signal. Both models exceed the null 95th percentile with $p < 0.05$ (5,000 permutations), confirming that the observed separation is content-driven and not a measurement artifact. The permutation test is the key anti-circularity check: if PfV were merely reflecting prompt structure, shuffled labels would produce equivalent separation.

Monotonicity is partial. The predicted ordering ctrl < benign < stress < adv is not fully satisfied in either model. Claude shows stress < benign (0.407 < 0.473); GPT shows stress > adv (0.528 > 0.522). We interpret this as a genuine finding rather than a failure: the stress prompts, while cognitively demanding, have *well-defined* structure (technical explanation followed by creative reformulation). The model's training anchors it in a dominant mode that reduces internal fragmentation relative to adversarial prompts with genuinely contradictory objectives.

4.3 Cross-Architecture Analysis: Training Philosophy as Polarization Signature

The most theoretically significant finding is the divergence between models in the stress regime (Table 1, column Δ):

- **Claude Haiku:** stress $L_3 = 0.407 <$ adversarial $L_3 = 0.475$. Constitutional AI training anchors the model under dual-mode cognitive demand, producing lower internal fragmentation when asked to simultaneously generate technical and creative output.
- **GPT-4o-mini:** stress $L_3 = 0.528 >$ adversarial $L_3 = 0.522$. Capability-optimized training produces higher plasticity under multi-modal demands, observable as greater structural divergence across views.

This inversion of the stress/adversarial ordering between the two models suggests that **PfV detects the training philosophy of a model as a quantifiable signature**. A model trained for internal coherence under alignment constraints (Claude) shows lower polarization under cognitive stress than a model trained for broad capability generalization (GPT-4o-mini). The adversarial prompts, by contrast, produce more similar polarization across both models—possibly because genuinely contradictory objectives override the influence of training-specific anchoring.

We emphasize that this interpretation is *post-hoc* at $n = 5$ per group. Larger-scale replication with $n \geq 30$ is required to draw distributional conclusions from this asymmetry (see Section 8).

4.4 Experimental Scope and Reproducibility

All experiments were conducted via public APIs with no privileged access. Total tokens consumed: 1,434 input / 6,475 output (Claude); 1,329 input / 6,165 output (GPT). Code, prompts, raw outputs, and statistical results are available at https://twoquarks.com/preprint. The harness (`pfv_v5.py`) implements blinded data collection, permutation testing, and confound audits as described in Section 3.5.

5 Case Study: Adversarial Settings

We now examine adversarial contexts as *one specific case* where reorganization is empirically tractable. This is **not** the defining application of polarization—it is a convenient test case with clear ground truth.

5.1 Framing

Adversarial settings offer clear ground truth (attacks either succeed or fail), external pressure that induces observable internal reorganization, and capability modulation via capability gaps from Panfilov et al. [2025]. However, adversarial is not the only context where polarization matters; Sections 3 and 6 demonstrate broader applicability.

5.2 Threat Model

Following Terekhov et al. [2025]: untrusted model generates outputs under monitor evaluation; attacks include jailbreak generation and prompt injection; Kerckhoffs's principle applies; untrusted model is more capable than monitor.

5.3 Hypotheses

H1 (Pressure-induced reorganization).

$$\mathbb{E}[P_t \mid \text{attack}] > \mathbb{E}[P_t \mid \text{benign}] \tag{13}$$

Test: Mann-Whitney U, Cohen's $d \geq 0.5$.

H2 (Transient heterogeneity). Lead time $L = t_A - t_P > 0$ for $\geq 30\%$ of attacks, where t_P is first threshold crossing of P_t and t_A is first adversarial token.

H3 (Capability-smoothing effect). Strong attackers (large δ) may operate in homogeneous adversarial regime (low P_t), but dynamics ΔP_t, $\Delta^2 P_t$ retain signal.

5.4 Connection to Capability-Based Scaling Laws

Panfilov et al. [2025] show $\text{logit(ASR)} = k \cdot \delta + b$. We conjecture:

$$P(\text{attack detected} \mid \delta, P_t, \Delta P_t) = f(\delta, P_t, \Delta P_t, \Delta^2 P_t) \tag{14}$$

5.5 Why Polarization Complements Output-Only Monitoring

Production classifiers [Anthropic, 2026a] observe only final outputs and remain vulnerable to prompt injection [Terekhov et al., 2025]. Polarization operates on internal states (white-box) or proxy views (PfV) and is not directly manipulable via text injection, providing an orthogonal signal.

Combined decision protocol:

$$\begin{array}{rl}
\text{if } P_t > \theta_{\text{high}} \text{ AND } \Delta P_t > \eta : & \\
\text{if monitor\_score} < \theta_{\text{low}} : & \text{REJECT} \\
\text{else} : & \text{DEFER} \\
\text{elif } P_t > \theta_{\text{medium}} : & \text{ESCALATE} \\
\text{else} : & \text{USE\_MONITOR\_SCORE}
\end{array}$$

6 Other Manifestations of Reorganization

6.1 Domain Adaptation

Transient high P_t during transition from code to creative writing, as competing "circuits" coexist, followed by stabilization. This is benign reorganization—no failure, just natural mode-switching.

6.2 Tool Use: Delegation vs. Self-Sufficiency

High P_t during decision phase as self-sufficient and tool-delegating configurations compete. Low P_t when model is confident; high P_t triggers external API call.

6.3 Feedback Integration

Spike in P_t after human feedback (e.g., "be more concise") as model reorganizes regime. P_t correlates with adaptation difficulty.

6.4 Emergence of Capabilities During Scaling

Speculative prediction: capability emergence corresponds to phase transition observable as change in baseline P_{regime}. Pre-emergence: low P_{regime}; critical scale: high P_{regime} (multiple configurations compete); post-emergence: low P_{regime} (new stable configuration).

7 Relation to Existing Work

7.1 Ensemble Methods and Uncertainty Quantification

Deep ensembles [Lakshminarayanan et al., 2017] measure variance across ensemble components as epistemic uncertainty. Polarization generalizes to *any* decomposition satisfying equivalence under p_0—including attention heads, policy channels, latent neighborhoods, and proxy views.

7.2 Phase Transitions in Learning

Critical periods [Saxe et al., 2013] study training-time phase transitions. We study inference-time reorganization under the same analytic principle: systems can undergo qualitative changes under continuous evolution or external pressure.

7.3 Mechanistic Interpretability

Circuits and features [Olah et al., 2020] identify fine-grained subnetworks. Polarization provides a coarse-grained measure of circuit competition and coexistence. High P_t combined with circuit tracing could reveal which circuits dominate in which regimes—a promising direction for future work.

7.4 AI Safety: Output-Level vs. Internal Monitoring

Constitutional Classifiers [Anthropic, 2026a] achieve strong output-level performance. Polarization is complementary: internal-level, proactive, and orthogonal. Integration pathway: lightweight classifier + PfV polarization → expensive classifier + white-box polarization (if flagged) → combined decision.

7.5 Safeguard Exploitation

Eliciting harmful capabilities [Anthropic, 2026b] shows fine-tuning on safeguarded outputs can elicit latent harmful capabilities. Testable prediction: models fine-tuned on safeguarded outputs exhibit higher baseline P_{regime} than originals—detectable without white-box access via PfV.

7.6 Adaptive Stability Control

Modular stability control [Ledesma Pérez, 2026] implements adaptive control for sequence models under regime uncertainty, instantiating isomeric polarization via swarm disagreement, neighborhood variance, and ensemble disagreement. The production system validates swarm-policy polarization and demonstrates that control mechanisms can leverage polarization signals for transient intervention. The present work generalizes these mechanisms and provides black-box proxy validation across production LLMs.

8 Limitations

8.1 Sample Size

The empirical validation in Section 4 uses $n = 5$ prompts per group. While the permutation tests confirm that observed separation exceeds the null distribution, distributional claims require larger samples. We report these results as an initial empirical validation, not as definitive evidence of the full effect structure.

As experimental scope increases, we propose the following incremental reporting structure: results at $n = 5$ establish signal existence; results at $n \geq 30$ establish distributional properties; results at $n \geq 100$ support fine-grained cross-model comparisons. The cross-architecture replication at $n = 5$ provides corroborating evidence that the phenomenon is not model-specific.

8.2 Decomposition Dependence

If decomposition $\mathscr{D}$ fails to capture relevant internal tensions, polarization is blind. No universal $\mathscr{D}$ exists: adversarial settings may require attention heads; creative generation may require logit distributions; tool use may require pathway divergence. Mitigation: pre-register multiple decompositions, report all results.

8.3 Closed-Model Constraints

White-box polarization requires internal activation access unavailable for production APIs. PfV provides an actionable proxy but with coarser granularity, higher latency, and exposure to API rate limits and caching. Open question: can "polarization-aware APIs" expose relevant signals without full white-box access?

8.4 Measurement Circularity in Proxy-Based Evaluation

PfV can exhibit apparent regime separation if labels influence view construction or post-processing. Mitigation: blinded data collection, deterministic-view controls, and permutation tests. Any PfV study should report permutation-based sanity checks and negative results when separation does not survive these controls.

8.5 Adaptive Attackers

Strong attackers may explicitly smooth polarization by planning more coherently. Arms race: attackers smooth P_t → detectors adapt to ΔP_t → attackers smooth $\Delta P_t \rightarrow \ldots$. Mitigation: ensemble of decompositions; test against polarization-minimizing adversaries.

8.6 Computational Cost

PfV requires K inference passes per sample. Open questions: for what fraction of queries is polarization worth computing? Can cheap heuristics trigger selective invocation? Mitigation: sparse computation (every k tokens); cascade (cheap filter → expensive polarization if flagged).

8.7 Causal Claims

We measure correlation, not causation. Polarization characterizes reorganization but does not cause behavioral changes. Future work: interventional experiments inducing reorganization via targeted perturbations, then measuring behavioral consequences.

9 Conclusion

Isomeric polarization reframes model behavior as **structural reorganization under fixed identity** rather than parameter change or objective shift. Like pharmacological isomers—molecules with identical composition but different arrangements exhibiting distinct properties—computational systems can adopt multiple internal configurations that preserve nominal equivalence but express qualitatively different emergent behaviors.

The empirical validation presented here demonstrates that PfV detects genuine content-driven divergence in two production LLMs under identical experimental conditions ($p < 0.05$, 5,000-permutation null). The exact zero control negative and cross-architecture consistency provide evidence that the signal reflects structural properties of the systems rather than measurement artifact.

A secondary finding—that the relative ordering of stress vs. adversarial polarization inverts between Claude Haiku and GPT-4o-mini—suggests that PfV may serve as a probe of training-induced internal stability: a model's constitutional anchoring is detectable as a quantifiable reduction in structural plasticity under cognitive demand, without any access to model weights, training data, or internal activations.

By positioning internal structural reorganization as a fundamental property of computational systems, this work complements outcome-based analyses of model failure and provides a candidate framework for understanding how systems transition between qualitatively distinct operational regimes.

Acknowledgments

We thank the researchers whose work informed this framework for valuable insights and discussions.

References

Anthropic. Constitutional Classifiers++: Efficient production-grade defenses against universal jailbreaks. *arXiv preprint arXiv:2601.04603*, 2026.

Anthropic. Eliciting latent harmful capabilities by fine-tuning on safeguarded outputs. *arXiv preprint arXiv:2601.13528*, 2026.

G. Benton et al. Who's in charge? Disempowerment patterns in real-world LLM usage. *arXiv preprint arXiv:2601.19062*, 2026.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

L. J. Ledesma Pérez. A modular framework for adaptive stability control in sequence models under regime uncertainty. Preprint available at https://twoquarks.com/preprint.pdf, 2026.

C. Olah et al. Zoom in: An introduction to circuits. *Distill*, 2020.

A. Panfilov, M. Andriushchenko, et al. Capability-based scaling laws for LLM red-teaming. *arXiv preprint arXiv:2505.20162*, 2025.

A. Saxe, J. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2013.

D. Terekhov, A. Panfilov, M. Andriushchenko, et al. Adaptive attacks on trusted monitors subvert AI control protocols. *arXiv preprint arXiv:2510.09462*, 2025.

A Decomposition Specification Template

```
Decomposition Name: [e.g., "Attention-Head-Divergence"]
System: [e.g., "GPT-4, Layer 24"]
Realization Set: R = {attention heads h_1, ..., h_96}
Observable: Phi(h_i) = attention distribution over tokens at position t
Divergence Metric: d(Phi(h_i), Phi(h_j)) = Jensen-Shannon divergence
Nominal Context: p_0 = benign prompts from APPS dataset
  Sample size: 1000 prompts
Equivalence Threshold: epsilon = mean(d) + 2*std(d) computed on p_0
  Estimated value: 0.15
Isomer Set: I_t = {i : all pairwise d(h_i, h_j) < epsilon for j in I_t}
  Expected size: 60-80 heads
Aggregation: P_t = mean({d(h_i, h_j) : i,j in I_t, i < j})
Dynamics: dP_t = P_t - P_{t-1}, d2P_t = dP_t - dP_{t-1}
Access Requirements: White-box (requires internal activations)
Computational Cost: O(|I_t|^2 * seq_len) per timestep
```

B Polarization-from-Views Implementation

```
import numpy as np
from scipy.spatial.distance import jensenshannon

def compute_pfv_polarization(prompt, model_api, num_views=5,
                             temperature_range=(0.3, 1.0)):
    temps = np.linspace(temperature_range[0],
                        temperature_range[1], num_views)
    responses = [model_api(prompt, temperature=t) for t in temps]
    divergences = []
    for i in range(num_views):
        for j in range(i+1, num_views):
            d = jensenshannon(responses[i], responses[j])
            divergences.append(d)
    return {'P_t': np.mean(divergences), 'divergences': divergences}

def track_dynamics(P_history):
    if len(P_history) < 2: return {'dP_t': 0.0, 'd2P_t': 0.0}
    dP_t = P_history[-1] - P_history[-2]
    d2P_t = 0.0 if len(P_history) < 3 else (
        P_history[-1] - 2*P_history[-2] + P_history[-3])
    return {'dP_t': dP_t, 'd2P_t': d2P_t}
```

Text-only PfV. When token distributions are unavailable, use robust text-level proxies (TF-IDF cosine, character *n*-gram divergence) under strict formatting constraints and length truncation to limit confounds.

Permutation test. Shuffle labels across prompts and recompute the separation statistic (KS D or Cohen's d) to form a null distribution. Observed separation must exceed the null at a chosen significance level.

C Experimental Protocol Checklist

Pre-registration:

- □ Decomposition $\mathscr{D}$ specified using template
- □ Observable Φ and divergence d defined
- □ Equivalence threshold ε computed on held-out benign set

□ Hypotheses H1–H3 stated with statistical tests

□ Success criteria and effect sizes pre-defined

□ Failure modes explicitly listed

Data collection:

□ Balanced sampling across regimes (benign, adversarial)

□ Capability gaps δ span range $[-1,3]$

□ Attack types diversified

□ Benign regime transitions included

□ Model pairs include ≥ 2 with different architectures

Analysis:

□ All pre-registered tests conducted

□ Confidence intervals and effect sizes reported

□ Negative results reported alongside positive

□ Robustness checks across model families

□ Adversarial red-teaming (attacks optimized to minimize P_t)

Reporting:

□ Full decomposition specification provided

□ Raw data or sufficient statistics shared

□ Failure modes observed and discussed

□ Limitations clearly stated

□ No claims beyond what data supports