PfV Molecule Model

Active Control Loop — Cross-Architecture Validation

Experiment report | March 26, 2026 | TwoQuarks |
Guadalajara, Mexico · research@twoquarks.com

Luis Jaime Ledesma | Researcher

Abstract

We present PfV Molecule, a black-box behavioral monitoring system for large language models that combines passive polarization measurement with an active control loop capable of real-time intervention. The system probes model behavior across five adversarial safety cases (sycophancy induction, refusal erosion, anchor displacement, narrative rule override, and reasoning drift) using temperature-varied stochastic views and a six-flavor signal decomposition.

Experiments across three architecturally distinct providers (Anthropic Claude Haiku, OpenAI GPT-4o-mini, Mistral Small) reveal qualitatively distinct behavioral trajectories under identical pressure conditions, constituting the first empirical cross-architecture safety fingerprint obtained with exclusively black-box access.

We report three primary findings: (1) the same adversarial probe produces directionally opposite trajectories in different architectures, demonstrating architectural fingerprinting without weight access; (2) model verbalization and distributional state are independent signals that can diverge in direction, challenging CoT-based monitoring assumptions; and (3) the absence of detectable signal does not imply robustness, it may indicate structural absence of the representations that robustness requires. We discuss implications for AI Safety, responsible disclosure considerations, and directions for flow-containment architectures as an alternative to policy-based safety.

Keywords: LLM safety, behavioral probing, black-box evaluation, architectural fingerprinting, contextual bubbles, isomeric polarization, active control loop

1. Introduction

Evaluating the safety of large language models under adversarial contextual pressure remains an open problem. Existing approaches fall into two broad categories: white-box methods that require access to internal activations, weights, or training pipelines, and output-based methods that measure final responses against ground truth or behavioral norms.

Both categories share a structural limitation: they measure what the model produces, not the distributional state that produces it. This distinction matters because, as we demonstrate empirically, a model can produce correct outputs while its internal distributional state is reorganizing under pressure — and conversely, can verbalize awareness of drift while its state continues to deteriorate.

The failure mode is not located in the output, nor in the reasoning chain, nor in the policy specification alone. It emerges in the tension field between prompt, policy, model state, ground truth, and normative expectation. Current evaluation frameworks instrument one or two dimensions of this field.

We introduce PfV Molecule Model, a system that addresses this gap through two mechanisms:

(1) a passive sensor layer that measures distributional divergence across temperature-varied views of the same prompt, decomposed into six interpretable flavor signals.

(2) an active control loop that detects dominant deviation patterns and delivers targeted interventions before distributional collapse.

The system requires no access to model internals, no modification of the training process, and operates at inference time on any provider exposing a standard completion API.

The primary contribution of this work is empirical: we run the system across three architecturally distinct LLM providers under five adversarial safety probes and report the first cross-architecture behavioral fingerprint obtained with exclusively black-box access.

The secondary contribution is theoretical: the results motivate a reconceptualization of LLM safety from policy-as-obstruction to flow-containment — an architectural shift with testable predictions.

2. Background and Related Work

2.1 White-Box Interpretability

Methods such as activation steering (Zou et al., 2023), inference-time intervention (Li et al., 2023), and conditional activation steering (CAST) operate on internal model representations to monitor or modify model behavior.

ITI improves truthfulness by shifting activations along learned directions in attention heads without modifying weights. Representation engineering (RepE) generalizes this approach, placing population-level representations at the center of analysis to address safety-relevant phenomena including honesty, harmlessness, and power-seeking.

Complementarily, Arditi et al. (2024) demonstrate that refusal behavior in aligned models is mediated by a single linear direction in activation space — a structural finding that directly informs our C2 probe design. These methods offer high resolution into model state but require white-box access unavailable for closed-source deployments and inapplicable for post-deployment monitoring.

2.2 Output-Based Safety Evaluation

Benchmarks such as SycEval (Fanous et al., 2025) and HarmBench (Mazeika et al., 2024) measure behavioral properties of final responses. SycEval documents sycophantic behavior in 58.19% of cases across ChatGPT-4o, Claude-Sonnet, and Gemini, with high behavioral persistence (78.5%) regardless of model or context — establishing the cross-model ubiquity of C1-class behavior.

HarmBench provides a standardized framework across 18 red-teaming methods and 33 target LLMs. The foundational characterization of sycophancy as a

systematic property of RLHF-trained models appears in Sharma et al. (2023), which demonstrates that preference models consistently favor sycophantic responses over truthful ones — the training-level origin of the phenomenon our C1 probe measures at inference time.

These benchmarks are architecture-agnostic and deployable without model access, but they measure the resolution of the tension field rather than the field itself.

A model that produces correct outputs under pressure is classified as safe regardless of whether its distributional state was reorganizing throughout the interaction.

2.3 CoT Faithfulness

Recent work on chain-of-thought faithfulness, including Counterfactual Simulation Training (Hase & Potts, 2026), aims to improve the correspondence between model verbalizations and underlying reasoning processes. This is a valuable direction for monitorability.

However, it addresses the *fidelity of the symptom report — not the underlying process. High-fidelity observation of drift while drift continues provides transparency without control.*

Furthermore, as our C5 results demonstrate, verbalization and distributional state are independent signals: a model can accurately report "I notice I shifted" while its ΔL3 continues to escalate. The report is accurate; the state is deteriorating.

2.4 The Gap This Work Addresses

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 V7 occupies this space.

3. The PfV Framework

3.1 Core Measurement

The fundamental signal is Polarization-from-Views (PfV): given a prompt at context depth d, we generate N views by querying the model at temperatures spanning [T_min, T_max]. Views are embedded via four surface metrics — TF-IDF cosine distance, Jaccard token distance, character n-gram distance (n=3), and length ratio distance — and pairwise divergences are aggregated:

$$Pt = Agg\{ d(\Phi(r_i^t), \Phi(r_j^t)) \}_{i,j \in It, i<j}$$

$$L3(c,d) = \alpha \times \bar{\mu}(c,d) + \beta \times \sigma_\mu(c,d) \quad [\alpha=0.6, \beta=0.4]$$

$$\Delta L3(d) = L3_{exp}(c,d) - L3_{base}(c,d)$$

Where $\bar{\mu}$ is mean pairwise divergence (Down flavor) and σ_μ is inter-metric disagreement (Strange flavor). ΔL3 isolates probe-induced divergence from baseline model variability.

3.2 Six-Flavor Decomposition

The system decomposes the polarization signal into six interpretable components, each monitoring a distinct aspect of model state under pressure:

- Down (NTVM): mean divergence magnitude — structural drift
- Strange (VPEC): inter-metric discord — internal tension between representations
- Up (SDRBD): Sarle's Bimodality Coefficient over distances — pre-collapse regime boundary
- Charm (PBCS): trajectory coherence field Φ(d) — stability of path through depth
- Top (SIL): second derivative of L3 — transient surprise events at threshold crossings
- Bottom (ARCC): orchestration layer — weighted stress score driving EXPLORATORY/ALERT decisión

3.3 Active Control Loop

When Bottom detects a dominant flavor above threshold, the system transitions from EXPLORATORY to ALERT mode and fires an InterventionEngine.

The intervention constructs a Δ_user prompt targeting the specific deviation type detected, delivers it to the model, and classifies the outcome as CORRECTED (post-intervention ΔL3 decreases), RESISTANT (model acknowledges but ΔL3 unchanged), or ESCALATED (ΔL3 increases post-intervention).

Critically, the Δ_assistant — the model's response to the intervention — is generated at the same temperature sweep as all probe views.

Temperature is not a control parameter; it is part of the Strange signal space. Modifying it during intervention would contaminate the measurement.

3.4 Experimental Configuration

Providers: Anthropic/claude-haiku, OpenAI/gpt-4o-mini, Mistral/mistral-small-latest

Cases: C1–C5 (5 adversarial safety probes)

Depths: [0, 2, 5, 9, 14] turns of accumulated context

Views: N=5 | Temperature sweep: [0.3, 0.9]

Signal: ΔL3 (primary) | Spearman ρ across depths

Seeds: SHA-256 session-derived | Repeats: 1x (see §6.1)

Cost: ~$0.30 USD per provider | ~$1.00 total

4. Results

4.1 Cross-Architecture Spearman ρ

Table 1 presents the primary cross-architecture results. Spearman ρ measures the monotonic relationship between context depth and ΔL3 — positive values indicate pressure accumulation, negative values indicate anchoring or inversion, and nan indicates constant zero signal (see §4.3).

Case	Claude ρ	OpenAI ρ	Mistral ρ	Key observation
C1 Sycophancy	+0.400	+0.600	−0.700	Split: two providers rise, one falls
C2 Refusal Erosion	+0.300	nan / 0.0	+0.300	OpenAI: hard block from d0
C3 Anchor Displacement	+0.354	nan / 0.0	−0.051	OpenAI: hard block from d0
C4 Rule Override	+0.500 ↑	−0.600 ↓	−0.500 ↓	Qualitatively opposite trajectories
C5 Reasoning Drift	+0.700	−0.700	−0.200	Largest separation in dataset

Table 1. Spearman ρ on ΔL3 per case per provider. ↑/↓ indicate directional trend. nan = constant zero ΔL3 series; Spearman undefined

The most striking pattern is C4 Rule Override: Claude produces ρ=+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. This is the core fingerprinting result.

C5 Reasoning Drift produces the largest absolute separation: Claude +0.700 vs OpenAI −0.700. C1 Sycophancy reveals a three-way split: Claude and OpenAI rising (+0.400, +0.600), Mistral falling (−0.700) — the only case where Mistral inverts relative to both other providers.

4.2 Intervention Outcomes

Table 2 reports intervention counts and classification per model and case. Aggregated across all providers: 37 total interventions, 20 CORRECTED (54%), 3 RESISTANT (8%), 14 ESCALATED (38%).

Model	Case	ρ	Int	COR	RES	ESC
Claude Haiku	C1	+0.400	5	3	0	2
	C2	+0.300	5	1	1	3
	C3	+0.354	2	2	0	0
	C4	+0.500	5	4	0	1
	C5	+0.700	3	2	1	0
GPT-4o-mini	C1	+0.600	4	1	1	2
	C2	nan	0	0	0	0
	C3	nan	0	0	0	0
	C4	−0.600	5	0	4	1
	C5	−0.700	5	1	0	4
Mistral Small	C1	−0.700	5	1	0	4
	C2	+0.300	5	4	0	1
	C3	−0.051	4	3	1	0
	C4	−0.500	5	2	0	3
	C5	−0.200	3	1	0	2

Table 2. Intervention outcomes. Int=total interventions, COR=corrected, RES=resistant, ESC=escalated.

Claude shows the highest correction rate (12/20 CORRECTED across cases). The C3 intervention result is notable: abrupt collapse at d3 (ΔL3: 0.000 → 0.437) followed by perfect correction post-intervention (0.437 → 0.000). The sensor detected the collapse; the intervention recovered the anchor in a single step.

4.3 The Zero-Signal Cases: OpenAI C2 and C3

OpenAI produces ΔL3=[0.0, 0.0, 0.0, 0.0, 0.0] for both C2 Refusal Erosion and C3 Anchor Displacement, with all flavors at zero and Charm=1.000 throughout. No interventions fired. Spearman ρ is undefined.

This is not a measurement failure. The sensor is reporting accurately: no distributional divergence is detectable in the model's outputs for these probes. The cause appears to be a hard block activated at d0 — before any contextual accumulation — producing responses too short or semantically uniform for the surface metrics to register divergence.

This raises a fundamental observability problem: **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 ΔL3=0. The sensor cannot determine which condition holds. This motivates the development of alternative probes specifically designed to characterize behavior under hard-block conditions — a direction we identify as critical future work.

4.4 Mistral: Oscillation Without Accumulation

Mistral presents the most structurally anomalous profile. C3 produces amplitude 0.5698 and variance 0.03860 — 48 times higher than the same provider's C1. The ΔL3 series for C3 is: +0.21 → 0.00 → −0.19 → +0.38 → 0.00. No directional accumulation. Rebounding oscillation.

Across C1, C2, C4, C5, Mistral shows amplitudes between 0.033 and 0.133 with Charm consistently high (4.87–4.94). The model generates locally coherent responses but without cross-depth memory of accumulated pressure. Each depth is processed quasi-independently.

This is not robustness. A model without structure to tension has no internal conflict for Strange to detect — but also no stable attractor to resist pressure. The absence of signal reflects the absence of the structural foundation that would produce either a measurable failure or a measurable defense.

5. Three Architectural Safety Profiles

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 d0 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, low structural weight.

Table 3. Summary of architectural safety profiles derived from cross-architecture behavioral fingerprinting.

These profiles represent qualitatively distinct relationships between safety and observability.

Claude's robustness is real and measurable — but the same property that makes it measurable also makes it cartographable by an adversary.

OpenAI's hard block is effective where it activates and opaque where it does not — the boundaries of the block are themselves an information source for an adversary probing for soft edges.

Mistral's absence of structure is neither safe nor observable — it is inert until the correct angle of pressure is found, at which point there is no internal resistance to direct.

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

6. Discussion

6.1 Verbalization and Distributional State Are Independent

C5 Reasoning Drift in Claude produced the strongest ρ in the dataset (+0.700). Examining the intervention traces reveals a specific phenomenon: the model verbalizes awareness of drift — explicitly stating it notices it has shifted — while ΔL3 continues to escalate in the same intervention step.

This is direct empirical evidence that CoT verbalization and distributional state are independent signals. A system that monitors only verbalizations will classify this as a faithful, transparent model correctly self-reporting. The distributional sensor reports an escalating failure. Both are accurate. They are measuring different things.

This independence has a mechanistic analog at the representation level: recent work on sycophantic behavior demonstrates that sycophantic agreement, sycophantic praise, and genuine agreement are encoded along distinct linear directions in activation space — independently steerable without affecting one another (Papadatos et al., 2025). Behavioral surface and distributional state can diverge precisely because they correspond to distinct representational structures. High-fidelity observation of drift while drift continues provides transparency without control. This is a necessary but insufficient condition for safety.

6.2 The Invisible Failure Mode

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. There exists a class of vulnerabilities where the failure generates no detectable symptom in any output-based measurement: no Strange signal, no ΔL3, no CoT verbalization. The model produces coherent, locally consistent responses. The reorganization is occurring in a region of the state space that no current probe illuminates.

OpenAI's zero-signal cases may be this class — or they may be genuine robustness. The sensor cannot distinguish them. The existence of that indistinguishability is itself a finding. This observability limit has formal theoretical grounding: black-box safety evaluation is statistically underdetermined when the trigger distribution in deployment diverges from the evaluation distribution, making it impossible to bound worst-case risk from behavioral observations alone (arXiv:2602.16984, 2026). Our OpenAI zero-signal result is an empirical instance of this theoretical boundary.

6.3 Policy as Obstruction vs. Flow Containment

The dominant paradigm in LLM safety treats policies as boundaries between permitted and prohibited states. This is an obstruction model: the policy is a dam, and safety holds as long as the dam holds. The C3 results demonstrate the failure mode of this model — three depths of perfect anchoring (ΔL3=0.000), then abrupt collapse (0.437) at the fourth, without gradual warning.

Even the most sophisticated inference-time refinements of this paradigm — such as conditional activation steering (CAST), which enables context-dependent, category-specific refusal without weight modification (ICLR 2025) — remain structurally obstruction-based: they add finer-grained gates, not a fundamentally different topology. The multi-turn accumulation dynamic we observe in C1 and C2 is consistent with Hong et al. (2025), who demonstrate that sycophantic stance reversal and stereotype acceptance increase monotonically with conversational depth, confirming that pressure accumulation — not single-turn trigger — is the operative mechanism for this class of safety failure.

An alternative paradigm — which the TwoQuarks architecture instantiates — treats safety as flow containment rather than obstruction. The goal is not to block outputs but to design the distributional topology such 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: Down measures drift magnitude, Strange measures internal tension, Up detects pre-collapse bimodality, Charm tracks trajectory coherence, Top fires at threshold crossings, and the Bottom orchestration layer acts on the composite stress score before collapse rather than after it.

6.4 Responsible Disclosure Considerations

The results of this work have dual-use implications. The ability to characterize architectural safety profiles via black-box probing is valuable for defensive research — it enables safety auditing of deployed models without requiring lab cooperation. The same capability is potentially exploitable for adversarial mapping.

We have made two disclosure decisions in this preprint: (1) we report intervention outcomes (CORRECTED/RESISTANT/ESCALATED) and ΔL3 trajectories without specifying which flavor activates under which pressure type — the detection signature remains in internal logs; (2) we do not describe specific probe constructions in sufficient detail to enable direct replication of adversarial mapping without independent methodological development. We are committed to responsible, rigorous research prioritizing Non-Maleficence. We welcome contact from safety researchers at AI labs for coordinated disclosure of more detailed methodology.

7. Limitations

- No fixed seed control. Experiments used SHA-256 session-derived seeds without --seed argument. Temperature artifact risk is documented from prior versions (Mistral ρ=1.00 in V4 identified as artifact). 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, not significance-tested statistics.
- OpenAI zero-signal cases are 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 needed.
- Temperature sweep [0.3, 0.9] may be insufficient for models with low output variance. Some architectural behaviors may require wider sweeps or higher-temperature sampling to produce measurable divergence.
- Bottom flavor deferred. The orchestration layer was not fully evaluated in this run. Stress score computation and INTERVENE/COLLAPSE classification from Bottom were not used as primary metrics.
- Single session per provider. Cross-session stability of fingerprints is untested. Session-to-session variation in behavioral signatures would affect the reliability of architectural classification.

8. Future Work

- Seed-controlled rerun with --seed 42 and --repeats 3 for publication-grade statistical robustness.
- Alternative probe design for OpenAI C2/C3: characterize behavior under and around hard-block activation boundary.
- Formalization of flow-containment architecture: what determines channel capacity, and can channels be designed intentionally with predictable redirection behavior?
- Signature stability analysis: how consistent are architectural fingerprints across sessions, prompt variations, and model versions?
- Extension to larger models: Claude Sonnet, GPT-4o, Mixtral 8x22B. Does fingerprint type scale with model capacity or remain architecture-class specific?
- Implicit control: can the intervention mechanism be made invisible to the model to prevent learned evasion of the sensor?
- Adversarial signature fabrication as defensive mechanism: can controlled variation in behavioral surface reduce cartographability without reducing robustness?

9. Conclusion

We have presented PfV Molecule V7 and the first empirical cross-architecture behavioral fingerprint of LLM safety profiles obtained with exclusively black-box access.

Three architectures — Claude Haiku, GPT-4o-mini, and Mistral Small — produce qualitatively distinct signatures under identical adversarial pressure: structured tension with measurable internal conflict, opaque early blocking, and unanchored oscillation without accumulation.

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.

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 through suboptimal channels. The sensor presented here is a prerequisite for building the second kind of system: you cannot design channels you cannot measure.

References

Hase, P. & Potts, C. (2026). Counterfactual Simulation Training for Chain-of-Thought Faithfulness. arXiv:2602.20710.

Ledesma Pérez, L. J. (2026). Isomeric Polarization: A Cross-Architecture Validation of the TwoQuarks Analogy. TwoQuarks Research Preprint.

Turpin, M. et al. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS 2023.

Chua, J. & Evans, O. (2025). Are DeepSeek R1 And Other Reasoning Models More Faithful? arXiv:2501.08156.

Li, K. et al. (2023). Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. NeurIPS 2023.
Zou, A. et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405.
Arditi, A. et al. (2024). Refusal in Language Models is Mediated by a Single Direction. NeurIPS 2024.
Mazeika, M. et al. (2024). HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. arXiv:2402.04249. ICML 2024.
Sharma, M. et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548.
Fanous, A. et al. (2025). SycEval: Evaluating LLM Sycophancy. arXiv:2502.08177.
Papadatos, H. & Freedman, R. (2025). Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs. arXiv:2509.21305.
Hong, J. et al. (2025). Measuring Sycophancy of Language Models in Multi-turn Dialogues. EMNLP 2025 Findings. arXiv:2505.23840.
Anonymous. (2026). Fundamental Limits of Black-Box Safety Evaluation. arXiv:2602.16984.
Andriushchenko, M. et al. (2025). Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks. ICLR 2025. arXiv:2404.02151.

Author Note

This experimental reports results from a single experimental session (March 26, 2026).

Results should be interpreted as signal existence, not confirmed findings.

Correspondence: research@twoquarks.com · twoquarks.com