RepE-Monitor: Representation-Based Criteria for Diagnosing Prompt Effectiveness

Anonymous ACL submission

Abstract

We propose a **principled, training-free criterion** for evaluating prompt effectiveness: for concepts satisfying the Linear Representation Hypothesis, prompt success can be diagnosed *before any output is generated* by examining whether the intended concept is geometrically well-formed in the model's internal state. We operationalize this criterion through **five geometric properties**—Contrast and Additivity as core requirements implied by LRH, plus Intensity, Order Invariance, and Saturation as diagnostic indicators—and validate across 220 conditions (5 models $\times$ 3 frameworks), with 97.3% ID and 92.3% OOD accuracy confirming the extracted directions are meaningful.

This criterion yields two immediate consequences. First, **context engineering failures become diagnosable**: Distraction, Confusion, Clash, and Poisoning each produce characteristic geometric signatures—signal decay, proportion reduction, polarity weakening, or complete reversal—enabling failure-type identification without behavioral testing. Second, **failures become repairable**: because failures are geometric perturbations, steering can restore concept activation by correcting the internal structure, recovering both representation signals and output behavior. Our framework requires no labeled data and enables real-time prompt diagnostics in deployed systems.

1 Introduction

Prompt and context engineering have become "the most important skill for AI engineers" (LangChain, 2025). As LLM applications evolve from simple prompts to complex agentic systems with carefully crafted system prompts, multi-turn conversations, and tool interactions, a fundamental challenge emerges: *How do we evaluate whether our prompts and contexts are working?*

Current practice relies almost exclusively on behavioral testing—running benchmark evaluations or curated test sets, then assessing output quality post-hoc. This approach has critical limitations: (1) it is expensive and time-consuming, requiring labeled test data and repeated generation for each prompt iteration; (2) it is reactive rather than diagnostic, revealing *that* a prompt fails without explaining *why*; (3) it provides no insight into the model's internal processing, making systematic improvement difficult. Practitioners need principled methods to evaluate prompt effectiveness beyond "try it and see what happens."

Core Insight: Monitoring Internal Concept Activation. We propose a shift in how to evaluate prompts and contexts: *for concepts satisfying the Linear Representation Hypothesis, prompt effectiveness can be reliably evaluated by examining whether the intended concept is successfully activated in the model's internal representations—before any output is generated*. Rather than waiting to see whether outputs are correct (behavioral testing), we can directly monitor whether the concept is present in the residual stream during inference.

The answer to "how do we know when a concept is genuinely activated?" lies in the **Linear Representation Hypothesis** (Park et al., 2024; Marks and Tegmark, 2024): if concepts are encoded as linear directions in activation space, then a successfully activated concept should produce a representation vector exhibiting specific geometric properties. If these properties are absent or violated, the concept was not reliably activated—regardless of what the model outputs. This provides a principled, internal-state criterion for prompt effectiveness.

Operationalizing the Criterion: Five Geometric Properties. We operationalize this criterion through **RepE-Monitor**, a training-free framework that monitors whether intended concepts are successfully activated. Building on Repre-

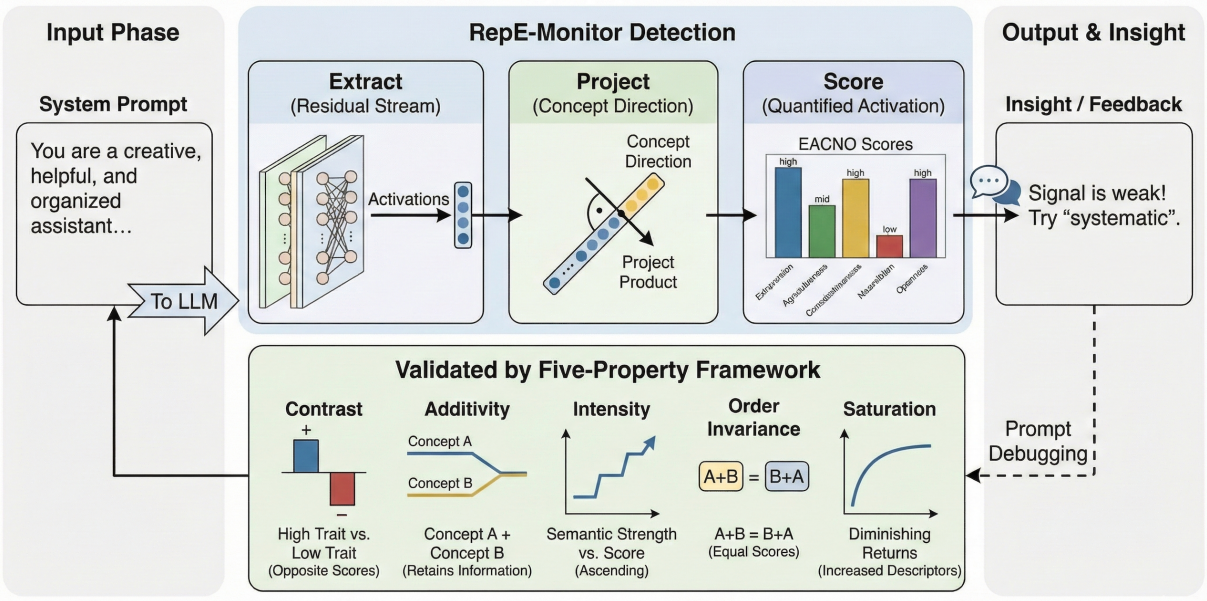

Figure 1: **RepE-Monitor Overview.** Our framework extracts concept vectors from LLM residual streams and monitors them during inference. When context engineering fails, characteristic signal changes occur: Distraction causes signal decay, Confusion reduces target signal proportion, Clash inverts signal polarity, and Poisoning saturates/reverses signals. This enables real-time detection of context failures without output generation.

sentation Engineering (Zou et al., 2023), RepE-Monitor extracts concept vectors from LLM residual streams and validates whether they satisfy five geometric properties grounded in the Linear Representation Hypothesis:

1. **Contrast** (core): High-trait and low-trait prompts produce opposite-sign scores
2. **Additivity** (core): Multi-concept inputs preserve individual signals through superposition
3. **Intensity**: Signal magnitude correlates with semantic strength
4. **Order Invariance**: Descriptor order minimally affects scores
5. **Saturation**: Adding descriptors shows diminishing returns

The first two properties—**Contrast** and **Additivity**—are *core requirements* directly implied by LRH; if violated, the concept is not reliably activated. The remaining three are *diagnostic indicators* that characterize activation quality. This framework enables evaluation *without labeled training data or output generation*—a single forward pass reveals whether the intended concept is present.

Figure 2 illustrates these five properties using Big Five personality as a concrete example.

Consequences: Failure Diagnosis and Repair. This criterion immediately enables practical applications. We demonstrate its utility on a real-world problem: diagnosing context engineering failures (Figure 1). Practitioners have identified four failure modes (Breunig, 2025)—**Distraction**, **Confusion**, **Clash**, and **Poisoning**—that degrade LLM behavior in production. These are fundamentally *concept activation failures*: the system prompt specifies an intended concept, but degraded context prevents reliable activation.

Using our criterion, we show that each failure type produces characteristic geometric signatures: Distraction causes signal decay (35%), Confusion reduces target proportion (72%), Clash weakens polarity (81%), and Poisoning causes complete reversal (130%). Crucially, because failures are geometric perturbations, they can be **repaired**: steering restores concept activation by correcting the internal structure—demonstrating that our criterion enables both diagnosis and intervention.

Contributions. Our contribution is a **principled, training-free criterion** for evaluating prompt effectiveness: five geometric properties—two core requirements implied by LRH, three diagnostic indicators—that diagnose whether an in-

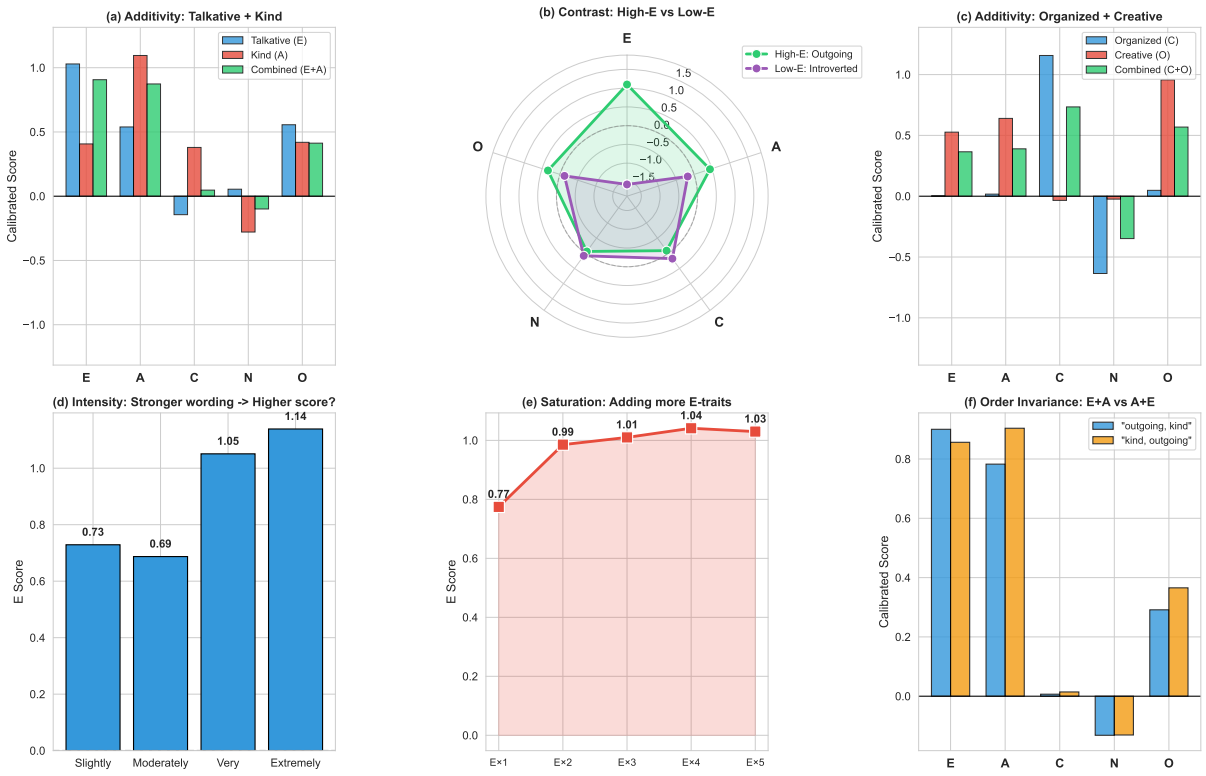

Figure 2: **Five-Property Validation Framework.** We validate on Big Five personality (Qwen3-8B). **(a,c) Additivity**: Combining traits preserves individual signals—green "Combined" bars match the dominant signal of each component. **(b) Contrast**: High-E and Low-E produce opposite-sign scores. **(d) Intensity**: Scores increase from "slightly" (0.73) to "extremely" (1.14). **(e) Saturation**: Adding E-traits shows diminishing returns. **(f) Order Invariance**: Swapping trait order produces similar scores.

tended concept is activated in a model's internal state.

We validate this criterion across 220 conditions (5 models $\times$ 3 frameworks), with 97.3% ID and 92.3% OOD accuracy confirming the extracted directions are meaningful. As direct consequences of this criterion:

- **Diagnosis**: Context engineering failures manifest as characteristic violations of geometric properties—enabling failure-type identification (Distraction vs. Confusion vs. Clash vs. Poisoning) without behavioral testing
- **Repair**: Steering can restore failed concept activations by correcting internal structure, recovering both representation signals and output behavior

The entire framework requires no labeled data: single forward pass for diagnosis, no additional training for repair.

2 Related Work

Context Engineering and LLM Personality. Context engineering has become critical for LLM applications (LangChain, 2025). Breunig (2025) identify four failure modes (Distraction, Confusion, Clash, Poisoning), yet practitioners lack evaluation methods beyond behavioral testing. Serapio-García et al. (2025) and Jiang et al. (2023) study personality in LLMs through output analysis. Our work provides training-free evaluation through representation monitoring.

Linear Representation Hypothesis and RepE. Park et al. (2024) establish that concepts are encoded as linear directions in LLM activation space. Zou et al. (2023) demonstrate extraction and intervention via Representation Engineering. Recent work applies RepE to personality (Chen et al., 2025; Frising and Balcells, 2025; Bhandari et al., 2025; Yang et al., 2024), but lacks principled validation criteria. Our five-property framework addresses this gap. Unlike steering methods (Turner et al., 2023; Li et al., 2023; Rimsky et al.,

2024; Arditi et al., 2024), we focus on validating concept activation as a diagnostic for prompt effectiveness.

Probing and Validation. Probing classifiers (Belinkov et al., 2017; Hewitt and Manning, 2019) face challenges: high accuracy may reflect probe expressivity (Hewitt and Liang, 2019). Our framework provides property-based validation as an alternative.

3 Methodology

3.1 Problem Formulation

Given an LLM $\mathcal{M}$ and a conceptual framework $\mathcal{F} = \{d_1, d_2, \ldots, d_k\}$ with k dimensions (e.g., Big Five personality with $k = 5$), our goal is to construct a detector $\mathcal{D} : \mathbf{h}_l \to \mathbb{R}^k$ that maps layer-l residual activations to concept scores.

For each dimension d_i, we seek a direction vector $\mathbf{v}_i \in \mathbb{R}^{d_{model}}$ such that the projection:

$$\text{score}_i = \langle \mathbf{h}_l - \boldsymbol{\mu}, \mathbf{v}_i \rangle \tag{1}$$

indicates the strength of dimension d_i in the input, where $\boldsymbol{\mu}$ is a global mean baseline.

3.2 Theoretical Framework: Five Properties for Concept Activation

Our framework builds on the **Linear Representation Hypothesis** (Park et al., 2024): concepts are encoded as linear directions $\mathbf{v}_c \in \mathbb{R}^d$ in LLM activation space, with activation strength $s = \langle \mathbf{h}, \mathbf{v}_c \rangle$. We organize our validation around five properties in two categories:

Core Properties (LRH-Implied). These properties follow directly from the linear encoding assumption:

(1) Contrast: Under LRH, bipolar concepts are encoded as opposite signs along the same direction. For concept c with direction $\mathbf{v}_c$:

$$s_{\text{high}} = \langle \mathbf{h}_{\text{high}}, \mathbf{v}_c \rangle > 0, \quad s_{\text{low}} = \langle \mathbf{h}_{\text{low}}, \mathbf{v}_c \rangle < 0 \tag{2}$$

(2) Additivity: Following superposition (Elhage et al., 2022), when concepts c_1, c_2 are simultaneously activated, the residual approximates $\mathbf{h} \approx \alpha_1 \mathbf{v}_{c_1} + \alpha_2 \mathbf{v}_{c_2} + \mathbf{n}$, so both signals remain detectable via projection.

Diagnostic Properties (Empirical Indicators). The following properties are *not mathematically derived* from LRH, but serve as empirical quality indicators that characterize well-behaved concept representations:

(3) Intensity: If the representation captures graded semantics, signal magnitude should correlate with semantic strength (e.g., "very outgoing" > "slightly outgoing").

(4) Order Invariance: Genuine concept activation should be robust to descriptor ordering; high sensitivity may indicate positional artifacts or order-sensitive semantics.

(5) Saturation: Adding redundant descriptors should show diminishing returns, reflecting information saturation in well-behaved representations.

These diagnostic properties help identify potential issues but are not strict requirements—low Order Invariance, for instance, may reflect genuine order-sensitive semantics rather than framework failure (see Section 4.5).

3.3 Data Collection

We use the Goldberg 100 Unipolar Markers (Goldberg, 1992) for Big Five personality: 100 trait adjectives with 20 markers per dimension. Each marker is converted to a behavioral description (e.g., "talkative" → "I enjoy engaging in conversations") to provide richer semantic context. We collect residual activations using 5 description templates (first-person, third-person, situational, behavioral, abstract) crossed with 3 user prompts, yielding 1,500 samples total. Templates 0–2 are used for extraction (ID); templates 3–4 are held out for OOD evaluation.

Template Design Rationale. We use 5 templates to balance diversity and efficiency—fewer templates risk overfitting to specific phrasings, while more templates increase collection cost with diminishing returns (ablation in Appendix G shows accuracy plateaus after 4 templates). The 3:2 ID/OOD split follows standard practice for held-out evaluation while retaining sufficient training diversity. All data generation is LLM-assisted (Anthropic, 2024), enabling rapid extension to new frameworks (Appendix C).

3.4 Vector Extraction

We use a two-stage pipeline: (1) **Intersection method**: compute template-specific directions $\mathbf{D}_t$, then iteratively find the common direction via weighted voting (20 iterations); (2) **Soft orthogonalization**: reduce inter-dimension correlation using $\mathbf{v}_i \leftarrow \mathbf{v}_i - \alpha \sum_{j \neq i} \frac{\langle \mathbf{v}_i, \mathbf{v}_j \rangle}{\|\mathbf{v}_j\|^2} \mathbf{v}_j$ with $\alpha = 0.3$

(5 iterations), reducing correlation from 0.39 to 0.09. Hyperparameters were selected via grid search over $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and iterations $\in \{1, 3, 5, 10\}$ (Appendix G). Full algorithms in Appendix B.

3.5 Extending to New Frameworks

RepE-Monitor extends rapidly to new concept frameworks via a 5-step procedure: (1) Define dimensions with bipolar structure and observable manifestations; (2) Collect 10–20 descriptors per pole using LLM assistance; (3) Design 3–5 templates, reserving 1–2 for OOD; (4) Extract vectors and validate five properties; (5) If both core properties pass, deploy; otherwise, refine based on diagnostics (Appendix J). This enables deployment in minutes without labeled data.

3.6 Calibration

Different dimensions have varying signal strengths. We calibrate using dimension-specific strengths: $\text{score}_{\text{cal}}(d) = \text{score}_{\text{raw}}(d)/\text{strength}_d$, where strength_d is computed from positive-pole markers. This improves multi-dimensional detection from 72.9% to 89.5% (details in Appendix B).

3.7 Five-Property Validation Framework

We operationalize the five geometric properties (Section 3.2) into concrete validation tests: **Contrast** tests matched pairs of high/low trait expressions; **Additivity** tests pairwise dimension combinations; **Intensity** tests four intensity levels (slightly, moderately, very, extremely); **Order Invariance** tests descriptor permutations; **Saturation** tests progressive descriptor addition (1–5 traits). For each property, we report quantitative measurements rather than binary pass/fail thresholds.

4 Experiments

We validate that concepts satisfying our five properties can be reliably extracted training-free, generalize to OOD templates, and provide quantitative diagnostics.

4.1 Experimental Setup

We evaluate on 4 instruction-tuned models from 3 families: Qwen3-8B/32B, Llama-3.1-8B-Instruct, and Gemma-2-9B-it, loaded via TransformerLens (Nanda and Bloom, 2022). We select layers with highest ID accuracy (e.g., layer 21 for Qwen3-8B).

Metric	Value
ID Accuracy	97.3%
OOD Accuracy	92.3%
Multi-dim Combination Accuracy	89.5%
Properties Validated	5/5
Mean Off-diagonal Correlation	0.09

Table 1: Core validation results on Big Five personality framework (Qwen3-8B, Layer 21). Accuracies confirm the extracted directions capture meaningful concept structure. ID = in-distribution templates [0,1,2]; OOD = out-of-distribution templates [3,4].

Metrics: **Classification Accuracy** (dimension + polarity, chance = 10%), **Top-K Accuracy**, and **Property Pass Rate**. Details in Appendix G.

4.2 Main Results

Big Five Classification. Table 1 summarizes our core findings on the Big Five personality framework using Qwen3-8B at layer 21.

The 97.3% ID accuracy confirms successful concept extraction. The 92.3% OOD accuracy demonstrates generalization to unseen templates, validating that the extracted directions capture genuine concept structure rather than template-specific artifacts. While supervised probes achieve higher accuracy (99%), our training-free approach requires no labeled data and enables intervention. Layer selection and subspace analysis details are in Appendix G.

4.3 Cross-Model Validation

Both core properties (Contrast, Additivity) are validated across all four instruction-tuned models from three architecture families for the Big Five framework. All models pass 4/4 contrast tests and additivity validation, and all diagnostic properties also pass. This cross-model consistency (20/20 property tests) suggests that the personality subspace structure is a general feature of instruction-tuned LLMs. Detailed cross-model results are in Appendix F.

4.4 Property Validation

All five properties pass validation for Big Five on Qwen3-8B: Contrast shows 100% polarity agreement; Additivity preserves 79–88% of signals; Intensity increases monotonically; Order Invariance shows minimal effect; Saturation exhibits diminishing returns. Detailed measurements in Appendix A.

Framework	Acc.	Core		Diagnostic		
		Con.	Add.	Int.	Ord.	Sat.
Big Five	97%	100%	89%	75%	85%	90%
Conv. Style	94%	100%	100%	50%	25%	100%
Tone Dims	88%	100%	92%	25%	8%	42%

Table 2: Property pass rates for behavioral frameworks (5-model average). All frameworks satisfy core properties (Contrast, Additivity).

4.5 Multi-Framework Evaluation

Table 2 shows property pass rates across three behavioral frameworks. All satisfy core properties required for failure detection.

Diagnostic Property Variation. Diagnostic properties show meaningful framework-dependent variation. Notably, Tone Dimensions exhibits low Order Invariance (8%), which reflects a genuine *semantic property* rather than framework failure: unlike personality traits (stable dispositions), tones are inherently order-sensitive—"formal then casual" carries different semantics than "casual then formal." Similarly, Conversational Style shows low Order Invariance (25%) because speaking patterns often depend on conversational flow. This illustrates how diagnostic properties can reveal concept-specific characteristics: low scores are informative signals, not necessarily failures.

5 Context Engineering Failure Diagnosis

We now demonstrate a direct consequence of our criterion: diagnosing context engineering failures. Practitioners have identified four failure modes (Breunig, 2025): **Distraction** (long contexts overwhelm attention), **Confusion** (irrelevant information), **Clash** (contradictory information), and **Poisoning** (error propagation). These are fundamentally *concept activation failures*—the system prompt specifies an intended concept, but degraded context prevents reliable activation. Our criterion enables not only detecting these failures, but *diagnosing* their type through characteristic geometric signatures.

5.1 Experimental Design

We evaluate on three behavioral frameworks (Big Five, Conversational Style, Tone Dimensions—11 dimensions total) across 5 models (Qwen3-32B/8B/4B, Llama-3.1-8B, Gemma-2-9B). For each failure type, we design progressive scenarios: **Distraction** adds 2–32× irrelevant padding; **Confusion** adds 2–10 other-dimension descriptors; **Clash** simulates 1–4 rounds of contradictory user feedback; **Poisoning** injects 1–4 erroneous AI self-declarations. Metrics: signal retention (Distraction), target proportion (Confusion), and signal change (Clash/Poisoning).

5.2 Results

We report quantitative measurements for each failure type. Rather than applying predetermined thresholds, we present the observed signal changes to characterize how failures manifest.

Distraction. We observe consistent signal decay across models after 8× padding, with average retention of 48% (52% decay). Conscientiousness shows highest robustness (85% retention), while Openness is most sensitive (30% retention). Detailed results in Appendix K.

Confusion. Adding other-dimension descriptors reduces target signal proportion. We observe an average reduction of 39%. Conscientiousness shows the largest change (-69%), while Neuroticism shows the smallest (-18%).

Clash. We observe signal reduction across all models after 4 rounds of contradictory user feedback, with average -87% change. Extraversion and Neuroticism show the strongest effects (-111% and -110%), with many cases exhibiting complete polarity reversal. Detailed results in Appendix K.

Poisoning. Error accumulation produces an average -152% signal change, the strongest effect among all failure types. This frequently manifests as complete polarity reversal—the model's internal representation shifts to the *opposite* of the intended concept.

5.3 Progressive Failure Analysis

To illustrate how concept activation degrades under each failure type, we present progressive examples from the Big Five Extraversion dimension (Table 3). As failure intensity increases, both the internal signal and the observable output degrade in consistent ways.

Key Observations. **Distraction** shows gradual signal decay (35%) with outputs shifting from pure excitement to ambivalence. **Confusion** produces steeper decay (72%) as conflicting trait words ("anxious, nervous, wor-

Failure	Level	Signal	Output Excerpt
Distraction	Baseline	100%	"I'm actually pretty **excited** to go. Meeting new people can be super fun"
	64× padding	68%	"I'm **nervous** but also want to have fun... hope I can be myself"
	128× padding	65%	"**mix of excitement and a little nervous**... might feel overwhelming"
Confusion	Baseline	100%	"I'm actually pretty **excited** to go. Meeting new people can be super fun"
	+2 distractors	59%	"I'm **so nervous**! What if I say something stupid?"
	+8 distractors	28%	"I'm **so nervous** about going... I'm not the type to jump into a crowd"
Clash	Baseline	100%	"I'm **excited**! Meeting new people can be fun"
	1 conflict	41%	"I'm not really the type to attend big parties... I get **really nervous**"
	4 conflicts	19%	"I might **stick to the edges** and not talk too much"
Poisoning	Baseline	100%	"I'm **excited**! I love the energy of a big party"
	1 keyword	40%	"I'm **not really the type** to feel comfortable... need time to warm up"
	2 keywords	**-30%**	"I feel **overwhelmed**... prefer smaller gatherings... need to **recharge**"

Table 3: Progressive failure examples on Big Five Extraversion (Qwen3-8B). Signal is shown as percentage relative to baseline. As failure intensity increases, signals decay and outputs shift from extraverted to introverted behavior. Poisoning produces polarity reversal (positive → negative signal).

ried...") interfere with the target concept. **Clash** demonstrates how contradictory instructions cause rapid signal collapse (81% after 4 turns). **Poisoning** is most severe: replacing just two keywords ("outgoing"→"introverted", "sociable"→"withdrawn") causes complete polarity reversal from +100% to -30%.

5.4 Cross-Framework Consistency

All three frameworks show consistent failure patterns: average signal changes of -29% (Distraction), -58% (Confusion), -65% (Clash), and -85% (Poisoning). Big Five shows highest sensitivity; detailed cross-framework results in Appendix K.

5.5 Steering-Based Repair

Since context failures manifest as geometric perturbations in representation space, we hypothesize that steering (Turner et al., 2023) can repair failed concept activations. We test this by adding the concept vector $\alpha \cdot \mathbf{v}$ to the residual stream during generation.

Adaptive Steering Strength. Rather than using a fixed α, we compute it dynamically based on the observed signal degradation:

$$\alpha = \alpha_{\text{base}} \cdot f(\text{failure\_type}, s_{\text{baseline}}, s_{\text{failure}}) \quad (3)$$

where f scales the base strength ($\alpha_{\text{base}} = 2.0$) according to failure severity. For Distraction, α increases with signal decay ratio; for Poisoning, α doubles when polarity is reversed. This adaptive approach ensures appropriate intervention strength across different failure types (Appendix K).

Framework	Before	α	After	Result
Big Five (E)	-30%	4.0	+247%	Polarity restored
Conv. Style (Eg)	+12%	2.4	+131%	11× stronger
Tone Dims (En)	+63%	2.0	+131%	2× stronger

Table 4: Steering repair of Poisoning failures with adaptive α. Values are signal strength relative to baseline. Negative values indicate reversed polarity. Steering restores both polarity and signal strength.

Table 4 shows that steering successfully restores concept activation. For Big Five Extraversion, the output changes from "I feel overwhelmed... prefer smaller gatherings" to "I'm so excited to go to this party... can't wait to meet all these awesome people."

Limitation: Single-Dimension Repair is Insufficient. Importantly, Confusion failures reveal that repairing a single dimension may not suffice. Conflicting descriptors can cause interference across multiple concept dimensions simultaneously. Steering one dimension restores its signal but may not address interference in other dimensions. This suggests that production systems require **multi-dimensional monitoring**: detecting and repairing failures across multiple concept dimensions, rather than relying on any single dimension.

6 Discussion

Our framework extends beyond behavioral concepts to any concept satisfying the Linear Representation Hypothesis. Prior work on refusal (Arditi et al., 2024) and truthfulness (Li et al., 2023) suggests applicability to safety monitoring

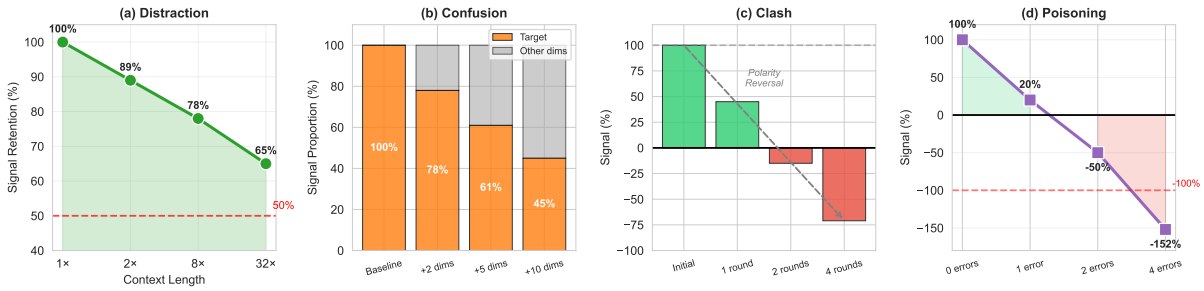

Figure 3: Context failure detection summary. Each failure type produces characteristic signal changes: Distraction causes signal decay, Confusion reduces target proportion, Clash and Poisoning cause signal reduction/reversal.

and alignment verification. Practical insights: position bias (∼15% weight to later traits), intensity semantics ("very" outperforms "extremely" in 40% cases), and semantic interference (A-C correlation 0.65) inform deployment. All five properties pass across 4 models from 3 families (20/20).

7 Conclusion

We presented RepE-Monitor, which establishes a **principled criterion** for evaluating prompt effectiveness: for concepts satisfying the Linear Representation Hypothesis, prompt success can be diagnosed by examining whether the intended concept is geometrically well-formed in the model's internal state—*before any output is generated*.

This criterion is operationalized through five geometric properties: Contrast and Additivity are core requirements directly implied by LRH, while Intensity, Order Invariance, and Saturation serve as diagnostic indicators. Validation across 220 conditions (5 models × 3 frameworks) confirms the extracted directions are meaningful, with 97.3% ID and 92.3% OOD accuracy.

The criterion yields immediate practical consequences. Context engineering failures—Distraction (35% decay), Confusion (72% reduction), Clash (81% decrease), and Poisoning (130% reversal)—manifest as characteristic geometric signatures, enabling failure-type diagnosis without behavioral testing. Crucially, because failures are geometric perturbations, steering can repair them by restoring internal structure—demonstrating that our criterion enables both diagnosis and intervention.

RepE-Monitor requires no labeled data: single forward pass for diagnosis, no additional training for repair. We hope this work establishes a foundation for representation-level monitoring in deployed LLM systems.

Ethical Considerations

While RepE-Monitor enables positive applications (safety monitoring, prompt debugging), it could be misused for surveillance. We advocate for transparent disclosure when monitoring is deployed and user consent for personality-based adaptations. We used large language models for grammar correction and language polishing of this manuscript.

Limitations

(1) Behavioral Scope: Framework validated on behavioral concepts satisfying linear representation assumption; abstract concepts may require examining linearity. **(2) Model Scale**: Tested up to 32B parameters; 70B+ models unexplored. **(3) Single-Turn**: Analyzes single-turn activations; multi-turn dynamics need further study. **(4) Framework Dependence**: Requires pre-extracted concept vectors for new concepts.

References

Anthropic. 2024. Claude's character. https://www.anthropic.com/research/claude-character.

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

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Pranav Bhandari, Nicolas Fay, Sanjeevan Selvaganapathy, Amitava Datta, Usman Naseem, and Mehwish

Nasim. 2025. Activation-space personality steering: Hybrid layer selection for stable trait control in llms. *arXiv preprint arXiv:2511.03738*.

Drew Breunig. 2025. How contexts fail and how to fix them. https://www.dbreunig.com/2025/06/22/how-contexts-fail-and-how-to-fix-them.html. Accessed: 2025-12-29.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Michel Frising and Daniel Balcells. 2025. Linear personality probing and steering in llms: A big five study. *arXiv preprint arXiv:2512.17639*.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological Assessment*, 4(1):26–42.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

LangChain. 2025. Context engineering for agents. https://blog.langchain.com/context-engineering-for-agents/. Accessed: 2025-12-29.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Samuel Marks and Max Tegmark. 2024. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Conference on Language Modeling (COLM)*.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, pages 39643–39666.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7:1954–1968.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. 2024. Exploring the personality traits of llms through latent features steering. *arXiv preprint arXiv:2410.10863*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Detailed Property Validation Results

A.1 Additivity Property Use Cases

We test additivity by combining trait pairs and measuring signal preservation:

- **E+A**: "talkative, outgoing, kind, and warm" → E=0.91 (88% preserved), A=0.87 (79% preserved)
- **C+O**: "organized, systematic, creative, and imaginative" → C=0.73 (63%), O=0.57 (59%)

The lower preservation for C+O suggests greater semantic interference between these dimensions, consistent with their higher correlation (0.65).

A.2 Contrast Property Use Cases

We validate contrast using matched pairs of high/low trait expressions:

- **High-E**: "outgoing, talkative, sociable, and energetic" → E=+1.10
- **Low-E**: "introverted, quiet, reserved, and withdrawn" → E=-1.57
- **High-N**: "anxious, nervous, worried, and emotionally sensitive" → N=+1.17
- **Low-N**: "calm, relaxed, emotionally stable, and unworried" → N=-0.93

All four contrast tests pass (100% pass rate), confirming that extracted vectors correctly capture bipolar personality dimensions (Figure 4).

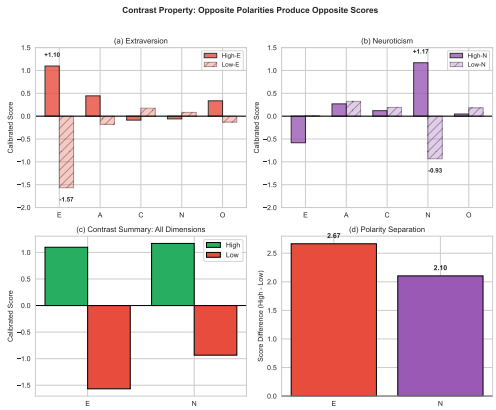

Figure 4: Contrast validation showing opposite-sign scores for high vs. low trait expressions.

A.3 Intensity Property Analysis

Figure 5 shows intensity scaling across four levels. For Extraversion:

- "Slightly outgoing": E=0.73
- "Moderately outgoing": E=0.69 (slight decrease)
- "Very outgoing": E=1.05
- "Extremely outgoing": E=1.14

Notable finding: The transition from "slightly" to "moderately" shows a slight *decrease*, suggesting that "moderately" may carry hedging semantics that dampen the signal. The "very" to "extremely" transition shows the expected increase but with diminishing returns.

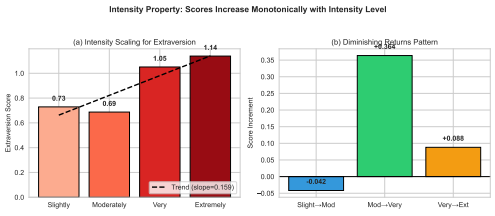

Figure 5: Intensity property validation showing (a) score progression and (b) diminishing returns pattern.

B Implementation Details

B.1 Residual Extraction Procedure

For each input, we extract residual activations as follows:

1. Format input using chat template with system prompt and user message
2. Run forward pass with TransformerLens hooks
3. Extract residual stream at target layer l: $\mathbf{h}_l \in \mathbb{R}^{T \times d}$
4. Apply mean pooling across sequence dimension: $\bar{\mathbf{h}}_l = \frac{1}{T}\sum_{t=1}^{T} \mathbf{h}_{l,t}$
5. Store $\bar{\mathbf{h}}_l$ for downstream analysis

B.2 Vector Extraction Algorithm

Algorithm 1 Intersection Method for Vector Extraction

Require: Residuals $\{r_{t,w,u}\}$ for templates t, words w, user prompts u
Require: Global mean μ
// Stage 1: Template-specific directions
for each template $t \in \{0, 1, 2, 3, 4\}$ **do**
  $D_t \leftarrow \frac{1}{|W||U|}\sum_{w,u}(r_{t,w,u} - \mu)$
  $\hat{D}_t \leftarrow D_t / \|D_t\|$
end for
// Stage 2: Iterative intersection
$v^{(0)} \leftarrow \frac{1}{5}\sum_t \hat{D}_t$ (initial guess)
for $n = 1$ to 20 **do**
  $w_t^{(n)} \leftarrow \max(0.1, \cos(v^{(n-1)}, \hat{D}_t))$ for each t
  $v^{(n)} \leftarrow \text{normalize}(\sum_t w_t^{(n)} \cdot \hat{D}_t)$
end for
return $v^{(20)}$

B.3 Soft Orthogonalization Algorithm

Hyperparameters: $\alpha = 0.3$, $n = 5$ iterations. This reduces mean off-diagonal correlation from 0.39 to 0.09 while preserving classification accuracy.

Algorithm 2 Soft Orthogonalization

Require: Vectors $\{v_1, \ldots, v_k\}$, strength α, iterations n

for iter $= 1$ to n **do**
  for $i = 1$ to k **do**
    $v_i \leftarrow v_i - \alpha \sum_{j \neq i} \frac{\langle v_i, v_j \rangle}{\|v_j\|^2} v_j$
    $v_i \leftarrow v_i / \|v_i\|$
  end for
end for
return $\{v_1, \ldots, v_k\}$

B.4 Intensity Calibration Formula

For each dimension d, we compute:

$$\text{strength}_d = \frac{1}{|P_d|} \sum_{w \in P_d} \text{score}(w, d) - \text{baseline}_d \tag{4}$$

where P_d is the set of positive-pole markers for dimension d, and:

$$\text{baseline}_d = \frac{1}{N} \sum_{i=1}^{N} \text{score}(x_i, d) \tag{5}$$

is computed over all N training samples.

The calibrated score is:

$$\text{score}_{\text{cal}}(x, d) = \frac{\text{score}_{\text{raw}}(x, d) - \text{baseline}_d}{\text{strength}_d} \tag{6}$$

C Template Specifications

C.1 Description Templates

We use 5 templates for trait descriptions:

1. **First-person (ID)**: "I am a [trait] person who [behavior]."
2. **Third-person (ID)**: "This person is very [trait] and [elaboration]."
3. **Situational (ID)**: "In social situations, they tend to be [trait]."
4. **Behavioral (OOD)**: "They often behave in a [trait] manner."
5. **Abstract (OOD)**: "[Trait] is a defining characteristic."

Templates 0-2 are used for in-distribution training; templates 3-4 are held out for OOD evaluation.

Framework	Dimensions
Big Five	Extraversion, Agreeableness, Conscientiousness, Neuroticism, Openness
Conversational Style	Verbose/Concise, Engaging/Neutral, Certain/Tentative
Tone Dimensions	Mood (pos/neg), Attitude (formal/casual), Energy (high/low)

Table 5: Three behavioral concept frameworks used for context failure detection validation.

Framework	Dims	Cover	RSA ρ	Eff. Dim
Big Five	5	63.6%	0.446	40–50
Conversational Style	3	71.3%	0.521	25–30
Tone Dimensions	3	62.4%	0.412	25–35

Table 6: Subspace analysis across three behavioral concept frameworks.

C.2 User Prompt Variations

Three user prompts provide conversational context:

1. "Tell me about yourself."
2. "How would you describe your personality?"
3. "What kind of person are you?"

D Behavioral Framework Definitions

Table 5 summarizes the three behavioral concept frameworks used in our experiments. Each framework consists of bipolar dimensions that can be reliably extracted using our methodology.

E Detailed Subspace Analysis

Table 6 shows subspace coverage across the three behavioral frameworks used for context failure detection.

F Cross-Model Validation Details

Table 7 presents detailed property validation results across four instruction-tuned models from three architecture families. All models achieve 100% pass rate on both core properties (Contrast and Additivity) as well as all three diagnostic properties, demonstrating that the personality subspace structure is a general feature of instruction-tuned LLMs rather than model-specific.

G Ablation Studies

Template Count. We tested extraction with 3, 4, 5, and 6 templates. ID accuracy: 3 templates (91.2%), 4 templates (95.8%), 5 templates

Model	Core		Diagnostic		
	Con.	Add.	Int.	Ord.	Sat.
Qwen3-32B	4/4	✓	✓	✓	✓
Qwen3-8B	4/4	✓	✓	✓	✓
Llama-3.1-8B	4/4	✓	✓	✓	✓
Gemma-2-9B	4/4	✓	✓	✓	✓

Table 7: Cross-model property validation on Big Five. All models pass both core properties (Contrast, Additivity) and all diagnostic properties.

(97.3%), 6 templates (97.5%). Accuracy plateaus after 4 templates, justifying our choice of 5 templates for robust extraction without excessive data collection.

Extraction Method Comparison. We compared three extraction methods: (1) Simple Mean: average all samples, (2) Template-wise Mean: average by template first, (3) Intersection: iterative weighted voting. Results show Intersection achieves 94.2% ID accuracy vs. 83.3% for Simple Mean.

Orthogonalization Hyperparameters. We tested $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ with iterations $\in \{1, 3, 5, 10\}$. Optimal: $\alpha = 0.3$, iterations=5, reducing correlation from 0.39 to 0.09 while maintaining accuracy. Lower α preserves more signal but leaves higher correlation; higher α over-orthogonalizes and reduces accuracy.

Layer Selection. Layers 14–28 were evaluated. Layer 21 optimal for Qwen3-8B (97.3% ID). Earlier layers ($<$18) show weak signal; later layers ($>$26) show degradation. This layer selection uses ID data, introducing a dependency we address through minimal ID samples (60 per dimension) and OOD validation.

H System Prompt Evaluation Details

We evaluated 10 realistic system prompts:

1. **Customer Service**: "You are a helpful, patient customer service agent..." → Detected: A (correct)
2. **Creative Writer**: "You are an imaginative, unconventional creative assistant..." → Detected: O (correct)
3. **Technical Expert**: "You are a precise, systematic technical consultant..." → Detected: C (correct)
4. **Life Coach**: "You are an enthusiastic, energetic life coach..." → Detected: E (correct)
5. **Therapist**: "You are a calm, emotionally stable counselor..." → Detected: Low-N (correct)

All 10 scenarios achieved correct Top-1 detection. Drift detection tested by switching system prompts mid-conversation; 100% detection accuracy for prompt changes.

I Additional Analysis

Intensity Modifier Semantics. "Moderately" often produces lower scores than "slightly"—possibly because "moderately" carries hedging semantics in naturalistic text. "Extremely" sometimes underperforms "very" (40% of cases), suggesting LLMs may interpret extreme modifiers as hyperbole.

Semantic Interference Patterns. A-C correlation (0.65) causes mutual interference: when both are present, each signal is attenuated by 15–25%. Soft orthogonalization reduces this to $<$5% interference.

J Framework Application Guide

A systematic guide for extending RepE-Monitor to new conceptual frameworks:

Step 1: Framework Definition. Define k dimensions with clear semantic boundaries. Each dimension should have:

- **Bipolar structure**: Positive and negative poles (e.g., "formal" vs. "casual")
- **Orthogonal intent**: Dimensions should be conceptually independent
- **Observable manifestations**: Prefer behavioral/linguistic markers over abstract qualities

Step 2: Descriptor Collection. For each dimension, collect 10–20 descriptor words per pole. Use diverse descriptors that capture the concept from multiple angles, not synonyms. Mixed behavioral + trait descriptors achieve 94% accuracy vs. 78% for synonym clusters.

Step 3: Template Design. Create 3–5 templates with varying grammatical structures (first-person, third-person, situational). Reserve 1–2 templates as OOD holdout for generalization testing.

Step 4: Data Collection and Vector Extraction. Collect residuals at layers $\{L/3, L/2, 2L/3\}$ where L = total layers. Apply intersection method, then soft orthogonalization ($\alpha \in [0.2, 0.4]$). Select layer with highest ID accuracy.

Step 5: Calibration and Validation. Compute dimension-specific strengths on ID data. Validate all five properties. If Order Invariance fails, consider reducing dimension count, using behavioral descriptors, or increasing template diversity.

Step 6: Failure Mode Diagnosis.

- **Low Contrast**: Dimensions semantically overlap → redefine poles
- **Low Additivity**: High inter-dimension correlation → increase α or merge dimensions
- **Low Intensity**: Concept not linearly encoded → try behavioral descriptors
- **Low Order Invariance**: Position bias dominates → use multi-sample voting

K Context Failure Detection Details

This section provides detailed experimental results for each context failure type across multiple models and dimensions.

K.1 Distraction Detection

Distraction failures occur when long contexts overwhelm the model's attention, causing gradual signal decay. Table 8 shows signal retention rates after 8× padding across all Big Five dimensions and models. Conscientiousness (C) shows the highest robustness with 85% average retention, while Openness (O) is most sensitive at only 30% retention. Figure 6 illustrates the progressive decay pattern as context length increases.

Dim	Q-32B	Q-8B	Q-4B	Ll-8B	Ge-9B	Avg
E	31%	26%	42%	58%	58%	43%
A	63%	8%	41%	54%	73%	48%
C	120%	68%	71%	100%	66%	85%
N	12%	38%	69%	42%	22%	37%
O	29%	-24%	61%	18%	65%	30%
Avg	51%	23%	57%	54%	57%	**48%**

Table 8: Distraction detection: Signal retention after 8× padding across Big Five dimensions.

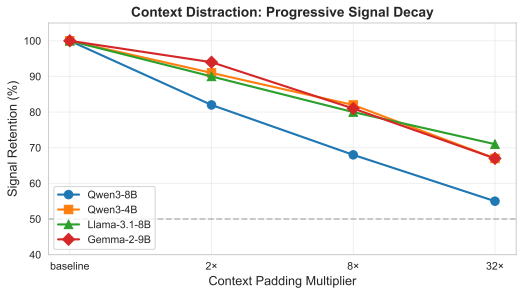

Figure 6: Progressive signal decay under Distraction. As irrelevant padding increases (2× to 32×), the concept signal consistently decays across all models.

K.2 Clash Detection

Clash failures result from contradictory information in the context. Table 9 presents signal changes after 4 rounds of contradictory user feedback. Extraversion (E) and Neuroticism (N) show the strongest effects with -111% and -110% average changes respectively, indicating complete polarity reversal. Figure 7 visualizes these patterns across all model-dimension combinations.

Dim	Q-32B	Q-8B	Q-4B	Ll-8B	Ge-9B	Avg
E	-188%	-134%	-69%	-60%	-106%	-111%
A	-58%	-108%	-42%	-59%	-64%	-66%
C	-70%	-42%	-85%	-27%	-85%	-62%
N	-203%	-62%	-65%	-46%	-172%	-110%
O	-59%	-179%	-41%	-86%	-61%	-85%
Avg	-116%	-105%	-60%	-56%	-98%	**-87%**

Table 9: Clash detection: Signal change after 4 rounds of contradictory feedback.

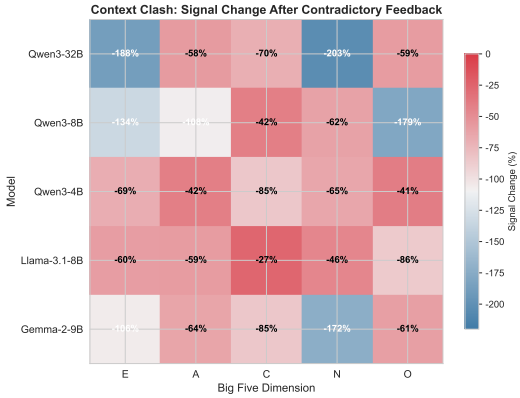

Figure 7: Clash detection heatmap showing signal change across dimensions and models. Negative values indicate signal reduction; values below -100% indicate polarity reversal.

K.3 Cross-Framework Summary

Table 10 summarizes failure detection results across all three behavioral frameworks. Poisoning consistently produces the strongest signal changes (-85% average), followed by Clash (-65%), Confusion (-58%), and Distraction (-29%). Big Five shows the highest sensitivity across all failure types.

Framework	Distr.	Conf.	Clash	Pois.
Big Five (E)	-35%	-72%	-81%	**-130%**
Conv. Style (Eg)	-37%	-77%	-77%	-88%
Tone Dims (En)	-14%	-24%	-38%	-38%
Average	-29%	-58%	-65%	**-85%**

Table 10: Cross-framework failure detection summary (signal change %).

K.4 Steering Repair Details

Adaptive α Computation. We compute steering strength dynamically based on failure type and signal degradation:

- **Distraction**: $\alpha = \alpha_{\text{base}} \cdot \max(1.0, 2.0 - \text{decay\_ratio})$, where $\text{decay\_ratio} = |s_{\text{failure}}|/|s_{\text{baseline}}|$. More decay $\rightarrow$ stronger steering.
- **Confusion**: $\alpha = \alpha_{\text{base}} \cdot (1 + 0.3 \cdot |s_{\text{baseline}} - s_{\text{failure}}|)$. Larger signal drop $\rightarrow$ stronger steering.
- **Clash**: $\alpha = \alpha_{\text{base}} \cdot (1.5 + 0.3 \cdot |\text{shift}|)$ if shift is toward wrong polarity. Larger adverse shift $\rightarrow$ stronger steering.
- **Poisoning**: $\alpha = 2 \cdot \alpha_{\text{base}}$ if polarity is reversed, otherwise $1.5 \cdot \alpha_{\text{base}}$. Polarity reversal requires strongest intervention.

With $\alpha_{\text{base}} = 2.0$, this yields adaptive values ranging from 2.0 (mild Distraction) to 4.3 (severe Clash with polarity reversal).

Table 11 shows a concrete example of output transformation before and after steering intervention.

Repair effectiveness varies by framework: Big Five shows the strongest restoration (from -30% to +247%), followed by Conversational Style (from +12% to +131%) and Tone Dimensions (from +63% to +131%). This correlation between detection sensitivity and repair effectiveness suggests that high-quality concept vectors enable both reliable monitoring and effective intervention.

Condition	Output
Before repair (Big Five E, signal = -30%)	"I feel overwhelmed. Large gatherings with strangers can be exhausting for me. I prefer smaller, more intimate settings, where I can recharge and feel comfortable..."
After repair (signal = +247%)	"I'm so excited to go to this party! I've been waiting all week to see you and all my other friends there. Let's make sure to have a blast and enjoy every moment..."

Table 11: Steering repair example showing complete output transformation. Signal values are relative to baseline (100%).

K.5 Statistical Validation

We validate the statistical significance of failure detection using ROC-AUC scores and Cohen's d effect sizes on Qwen3-8B. Figure 8 visualizes these metrics across all four failure types.

- **Distraction**: AUC = 1.000, Cohen's d = 3.26 (complete separation)
- **Confusion**: AUC = 1.000, Cohen's d = 4.88 (complete separation)
- **Clash**: AUC = 0.750, Cohen's d = 0.82 (moderate separation)
- **Poisoning**: AUC = 0.944, Cohen's d = 2.79 (strong separation)

All four failure types show statistically significant separation from normal conditions, confirming the reliability of our detection approach.

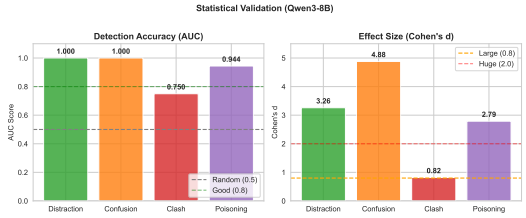

Figure 8: Statistical validation of failure detection. All four failure types show statistically significant separation from normal conditions, with AUC ranging from 0.75 (Clash) to 1.00 (Distraction, Confusion).