

# Explore-Consolidate Dynamics in Cross-Probe Coherence Separate Successful and Failed LLM Agent Trajectories

Caio Vicentino (ORCID: 0009-0003-4331-6259) OpenInterpretability — caio@openinterp.org

**Draft v2.** 2026-05-18. Target venue: NeurIPS MI Workshop 2026 (Sep submission) or ICLR 2027 main track.

---

## Abstract

Behavioral probes for LLM monitoring are typically evaluated one axis at a time. We propose a meta-signal — *cross-probe coherence*  $\kappa_t$  — measured as the mean absolute pairwise correlation across  $N$  concurrent per-turn probes within a moving window of agent turns. On 99 SWE-bench Pro trajectories from Qwen3.6-27B (3 872 turns, 5 per-turn probes trained on a single layer of the residual stream), we report two distinct findings. (1) Per-trace mean  $\bar{\kappa}$  separates success from failure at AUROC 0.677 (Mann-Whitney  $p=0.0009$ ). (2)  $\kappa_t$  exhibits a **U-shape** over each trajectory: it decreases through the first half (exploration) and increases through the second half (consolidation), and the *amplitude* of this U-shape is markedly larger in successful traces. The early-half slope is  $-0.0078/\text{turn}$  in success vs.  $-0.0007/\text{turn}$  in failure ( $p=0.0002$ ); the late-half slope is  $+0.0149/\text{turn}$  in success vs.  $+0.0025/\text{turn}$  in failure ( $p=0.00004$ ). A pre-registered robustness control (C1) found the monolithic per-trace slope is largely explained by trace-length confound ( $p=0.56$  after length regression), motivating the length-normalized early-half/late-half decomposition. Within-trace turn-order shuffle nulls (C2) confirm the U-shape is genuinely temporal ( $p<0.0001$ ). The pattern is the *inverse* of cardiac uncoupling: in ICU literature, cross-vital decorrelation anticipates physiological decompensation; in LLM agents, cross-probe trajectories *oscillate* during successful reasoning (explore → consolidate) and remain flat during failure. We discuss the methodological discipline — pre-registered gates that caught both five prior single-probe walk-backs and this paper’s own headline-confound — that gives the finding its credibility.

---

## 1. Introduction

### 1.1 Single-probe monitoring is the default

Production LLM monitoring stacks — Apollo’s Watcher series, Anthropic’s “Cheap Monitors” and Persona Vector probes, OpenAI’s internal classifiers — typically deploy *one probe per failure mode*: one deception classifier, one sycophancy classifier, one capability classifier. When multiple probes are available, they are evaluated in isolation or combined by simple aggregation (averaging logits, OR-gating on thresholds; see Hua et al., 2025).

This pattern is sensible: each probe is its own scientific artifact, with its own training distribution, calibration, and failure modes. But it implicitly assumes that any signal in the *joint distribution* of probe outputs is already captured by the marginals. We argue — and empirically demonstrate — that this assumption can be false. The *covariance structure* across  $N$  concurrent probes carries meta-signal that no single probe captures alone.

### 1.2 Biological inspiration: cardiac uncoupling

Twenty years of intensive-care literature, beginning with Morris (2006), documents that healthy patients show high *cross-vital coherence*: heart rate, blood pressure, and respiration rate are strongly coupled through autonomic feedback. Patients on a trajectory toward physiological decompensation show **cross-vital uncoupling** — the correlations collapse — *hours* before the crisis becomes overt. Modern Early

Warning Scores deployed across tens of thousands of patients use this pattern as a validated mortality predictor.

The structural similarity to multi-probe monitoring of LLM agents is direct. Each probe is a “vital sign” measured at each turn. The question of whether their cross-coupling carries predictive information about reasoning success or failure is empirically testable.

### 1.3 What we contribute

1. We define cross-probe coherence  $\kappa_t$  for  $N$  concurrent behavioral probes evaluated per agent turn, on a moving window of  $W$  turns.
2. We show that per-trace mean  $\bar{\kappa}$  separates SWE-bench Pro success from failure at AUROC 0.677 (lift +0.176 over a shuffled baseline; Mann-Whitney  $p=0.0009$ ).
3. We characterize the *temporal shape* of  $\kappa_t$  over agent trajectories as a **U**: a falling early phase (“exploration”) and a rising late phase (“consolidation”). The amplitude of this U is substantially larger in successful trajectories than in failed ones: early-half slope  $-0.0078/\text{turn}$  (success) vs.  $-0.0007/\text{turn}$  (failure),  $p=0.0002$ ; late-half slope  $+0.0149/\text{turn}$  (success) vs.  $+0.0025/\text{turn}$  (failure),  $p=0.00004$ .
4. We document and resolve a length-confound: a naïve per-trace slope (averaging the early-fall and late-rise) is explained by trace-length variation across classes (regression-residual Mann-Whitney  $p=0.56$ ). The length-normalized early/late half decomposition is the correct primary statistic; the monolithic slope was a misleading summary.
5. We show the pattern is *inverse* to its biological analogue: cardiac decompensation involves cross-channel **collapse** from coupled baseline; LLM-agent failure involves cross-probe **flatness**, the absence of explore → consolidate oscillation rather than a coupling collapse.
6. We document the methodological discipline — five pre-registered single-probe candidates walked back in the 36 hours before this finding, plus this paper’s own pre-registered controls catching a headline-confound — as the load-bearing source of credibility.

### 1.4 Methodological context

This finding is the sixth probe project in a single 36-hour sprint. The five preceding candidates — silent-refusal (irreproducible across transformers commits), tool-doubt (lexical leak via `ok=false` substrings), tool-doubt position-sweep (no semantic signal at any layer/position), Inner-Outer paired probe (capability redux), and  $\kappa_t$  v1 with trace-level labels (no within-trace dynamics) — were each governed by a four-gate pre-registration protocol that they failed. The present work passes the original four gates (G1, G2, G3 on  $\bar{\kappa}$ ; G4 marginal) but the bonus per-trace slope statistic, when subjected to a separately pre-registered five-control battery, was found to be substantially confounded with trace length (C1,  $p=0.56$ ). The correct length-normalized statistic — the early-half/late-half decomposition — was constructed *after* the controls flagged the confound, and replicates the substantive finding (succ vs fail differ significantly in both halves, in opposite directions). We narrate this walk-back-and-rescue trajectory explicitly because we believe the pre-registration discipline is what carries the credibility of the rescued finding.

## 2. Method

### 2.1 Cross-probe coherence $\kappa_t$

Let  $P_1, \dots, P_N$  be  $N$  behavioral probes, each producing a scalar score  $s_i(t)$  for a given agent turn  $t$  in a trace  $\tau$ . For a window half-width  $W$  (we use  $W=3$ , so a 7-turn window), define the windowed coherence at turn  $t$  as:

$$\kappa_t(\tau, t; W) = (1 / (N \cdot (N-1))) \cdot \sum_{i < j} |\text{corr}_\tau(s_i, s_j) \text{ over } t' \in [t-W, t+W]|$$

Where  $\text{corr}_\tau$  is the Pearson correlation of probe scores computed only over the turns in the window of trace  $\tau$ . Absolute value is used because we care about *coupling magnitude*, not sign; individual probes

may be encoded with opposite polarity.

Per-trace summary statistics used in this paper:

- **Mean coherence:**  $\bar{\kappa}(\tau)$  = mean of  $\kappa_t(\tau, t; W)$  over all turns  $t$  with a non-degenerate window.
- **Early-half slope:** linear regression of  $\kappa_t(\tau, t; W)$  against  $t$  over the first  $\lfloor n_{\text{turns}}(\tau) / 2 \rfloor$  turns. Denote  $\hat{s}_{\text{early}}(\tau)$ .
- **Late-half slope:** linear regression of  $\kappa_t(\tau, t; W)$  against  $t$  over the second half of  $\tau$ . Denote  $\hat{s}_{\text{late}}(\tau)$ .
- **Monolithic slope** (deprecated as primary statistic, retained for honesty): regression over all valid windows in  $\tau$ . Denote  $\hat{s}(\tau)$ .

The early/late decomposition is *length-normalized by construction*: each trace contributes one slope value per half regardless of its absolute length. This is the key technical move that survives the trace-length confound discussed in §4.

## 2.2 Probe stack

We train 5 per-turn probes on the SWE-bench Pro Phase 6 capture set: 99 traces  $\times$  ~40 turns/trace  $\times$  5 layers  $\times$  4 positions of saved residual stream (Vicentino, 2026b). All five probes are linear logistic regression with `StandardScaler` and `liblinear` solver ( $C=0.1$ , class-weighted), trained on the *paired concatenation* of `think_end` and `turn_end` residuals at layer 43 (concat dim 10 240). Cross-validation is `GroupKFold` by `instance_id` to prevent trace-level leakage.

Probe	Per-turn label	Positive rate	AUROC
A. tool_finish	turn calls finish?	0.9%	1.000
B. tool_bash	turn calls bash?	74.6%	1.000
C. long_thinking	thinking chars > trace median?	42.9%	0.990
D. tool_ok	tool_results[0].ok == True?	19.6%	0.998
E. repo_ansible	trace is in ansible repo?	32.2%	0.970

Pairwise probe-score correlations (Pearson, computed on out-of-fold scores): the strongest off-diagonal is B  $\square$  D at  $-0.83$  (bash tool calls fail at elevated rate); A  $\square$  D at  $+0.33$  (the `finish` tool trivially succeeds); C and E are essentially orthogonal to others ( $|\text{corr}| \leq 0.07$ ). Effective independent axes are approximately 3–4.

## 2.3 Pre-registered gates and controls

**Original four gates** (pre-registered before the v2 finding emerged, tied to project notes timestamped 2026-05-18 11:00 UTC; the same protocol governed the five walked-back siblings):

Gate	Threshold	Rationale
G1	$\bar{\kappa}$ AUROC > 0.65	Trace-level discrimination
G2	$\bar{\kappa}$ AUROC – shuffled AUROC > 0.10	Distinct from sampling artifact
G3	Mann-Whitney $p < 0.05$ ( $\bar{\kappa}$ across classes)	Distributions actually differ
G4	EARLY $\kappa_t$ (first 5 turns) AUROC > 0.60	True anticipation, not retrospection

Results: G1, G2, G3 pass; G4 marginal ( $0.572 < 0.60$ ). The mean- $\bar{\kappa}$  findings (§3.1) are pre-registered.

**Five robustness controls** (separately pre-registered after the bonus per-trace slope statistic was observed, before the controls were computed):

Control	What it tests
C1	Does total trace length explain per-trace slope?
C2	Does within-trace turn-order shuffle erase the class difference in slope?
C3	Does the slope effect survive restriction to maximally probe-decorrelated axes?
C4	(information) Does trace length differ by class?
C5	Does the effect hold within a fixed-length (max-turns-only) stratum?

Results are presented in §4; the substantive consequence is that the **monolithic slope statistic is confounded with trace length (C1,  $p=0.56$ )**, motivating the early-half/late-half decomposition (§2.1) as the primary post-control statistic.

## 2.4 Walk-back history (v1, then v2, then slope deprecation)

A v1 of  $\kappa_t$  used trace-level outcome labels for each per-turn probe; the probes collapsed to trace-constants and v1’s G4 EARLY  $\kappa_t$  was 0.458 (worse than chance). We rebuilt with per-turn labels in v2; G4 improved to 0.572 and the bonus slope statistic appeared to be the headline ( $p=0.0003$ ).

Subsequent controls then revealed that the v2 slope statistic was length-confounded. The early/late half decomposition was *not* in the original v2 pre-registration; it was added once C1 flagged the confound. It is therefore reported as a post-hoc rescue rather than as a pre-registered primary statistic. The mean- $\bar{\kappa}$  statistic of §3.1, by contrast, is the original pre-registered finding and is unaffected by the C1 finding.

## 2.5 Data

99 SWE-bench Pro instances, randomly sampled from the train split, capped at `max_turns=30` per trace. Qwen3.6-27B with `enable_thinking=True`, transformers commit pinned to 73d9159.... 11 reasoning-relevant layers captured (L3, L7, L11, L15, L23, L31, L43, L47, L51, L55, L59); 4 token positions per turn. All  $\kappa_t$  analysis runs on a 2024 MacBook Pro CPU; no GPU is needed beyond the initial Phase 6 capture run.

# 3. Results

## 3.1 Per-trace mean $\bar{\kappa}$ separates outcomes (pre-registered headline)

Per-trace mean  $\bar{\kappa}$  across the 99 traces:

	N	mean $\bar{\kappa}$	std
Success	40	0.437	0.069
Failure	59	0.417	0.071

Mann-Whitney U on  $\bar{\kappa}$  across classes:  **$p = 0.0009$** . AUROC of  $(-\bar{\kappa})$  as a failure predictor: 0.677; shuffled-label baseline AUROC: 0.501; gap +0.176. G1 ( $>0.65$ ), G2 ( $>0.10$ ), and G3 ( $<0.05$ ) all pass.

## 3.2 $\kappa_t$ trajectories show a U-shape (post-control headline)

Per-trace plots of  $\kappa_t$  over turns reveal a robust pattern:  $\kappa_t$  typically *decreases* through the first half of a trajectory and *increases* through the second half. We decompose each trace’s slope into an early half (first  $\lceil n\_turns/2 \rceil$  turns) and a late half (remainder), and test whether either half differs by outcome class:

	N	Mean early-half slope	Mean late-half slope
Success	39	<b>-0.0078</b> $\pm$ 0.0090	<b>+0.0149</b> $\pm$ 0.0094
Failure	59	-0.0007 $\pm$ 0.0034	+0.0025 $\pm$ 0.0046
Mann-Whitney p		<b>0.0002</b>	<b>0.00004</b>

Both halves significantly differentiate outcomes, but **in opposite directions**:

- In the early half,  $\kappa_t$  *decreases faster* in successful traces (more pronounced exploration / decoupling phase).
- In the late half,  $\kappa_t$  *increases faster* in successful traces (more pronounced consolidation phase).

Successful traces have a high-amplitude U-shape; failed traces have a flat trajectory. This is the substantive finding of the paper.

### 3.3 EARLY $\kappa_t$ (G4 absolute-value) is marginal

EARLY  $\bar{\kappa}$  (mean of  $\kappa_t$  over the first 5 turns, then AUROC against outcome): 0.572, below the pre-registered 0.60 threshold. G4 **fails**.

This is consistent with §3.2: the discriminative signal in the early phase is not the absolute *value* of  $\kappa_t$  but the *rate of change* (slope). The class difference in early-half slope (-0.0078 vs -0.0007) is starkly significant ( $p=0.0002$ ) while the difference in early-half absolute  $\bar{\kappa}$  is not.

### 3.4 Internal replication: 9 axes preserve the mean discrimination (v3)

We re-ran the  $\bar{\kappa}$  pipeline with an expanded probe set: 5 original axes plus 4 additional per-turn axes (F. response\_long; G. early\_trace; H. late\_trace; I. slow\_turn). All trained on the same L43 paired residual.

Metric	v2 (5 axes)	v3 (9 axes)	Trend
$\bar{\kappa}$ AUROC	0.677	0.697	strengthens
Gap vs shuffled	+0.176	+0.194	strengthens
Mann-Whitney p ( $\bar{\kappa}$ )	0.003	0.0009	3× sharper
Monolithic slope p	0.0003	0.003	weaker
Success monolithic slope	+0.0037	+0.0025	shrinks
Failure monolithic slope	+0.0003	+0.0002	similar

The trace-level mean discrimination strengthens. The deprecated monolithic-slope statistic weakens (in addition to being later found length-confounded). The early/late half decomposition was not yet computed at the time of the v3 run.

## 4. Robustness Controls

Five controls were pre-registered after the v2 finding and before they were computed. Numbers are from kappa\_t\_controls\_results.json.

### 4.1 C1 — Trace-length confound (FAIL; this finding is what motivated the U-shape decomposition)

**Hypothesis under test:** the per-trace monolithic slope  $\hat{s}$  might be inflated mechanically in shorter traces (shorter traces concentrate more of  $\kappa_t$ 's typical trajectory into fewer turns, biasing slope estimates). If so, the class difference in the monolithic slope might be a length difference in disguise.

**Test:** Regress per-trace slopes on per-trace  $n\_turns$  (linear), then test whether residual slopes differ by class via Mann-Whitney U.

**Result:**

- Pearson( $n\_turns$ , slope) =  $-0.41$ ,  $p = 2.1 \times 10^{-5}$
- Spearman( $n\_turns$ , slope) =  $-0.31$ ,  $p = 0.0017$
- Length-residualized slope: success residual mean  $-0.0001$ , failure residual mean  $+0.0001$
- Mann-Whitney  $p$  (residualized slopes) = **0.56**

**Verdict: the monolithic per-trace slope statistic is substantially confounded with trace length.** Length explains the class difference: once you remove the linear contribution of length, the class effect is gone.

We treat this as a *successful* control — the protocol caught a confound in a bonus statistic before publication. It motivated the early/late half decomposition reported in §3.2, which is length-normalized by construction.

#### 4.2 C2 — Within-trace turn-order shuffle (PASS)

**Hypothesis under test:** if the slope difference were *truly* temporal — buildup over the course of a trajectory — then permuting the order of turns within each trace should erase it. If the class difference survived permutation, then we would not be measuring temporal dynamics; we would be measuring something close to a class-conditioned mean.

**Test:** For each of 200 permutations, randomly shuffle each trace’s turn order and recompute per-trace slopes. Compute the success-vs-failure mean slope difference under the null.

**Result:**

- Real class difference (succ – fail) of monolithic slope:  $+0.0023$
- Null distribution under shuffle: mean =  $+0.00005$ , std =  $0.00070$
- Shuffle-permutation  $p$ : **<0.0001** (none of 200 permutations exceeded the real value)

**Verdict: temporal ordering matters.** The  $\kappa\_t$  signal is genuinely sequential, not just a class-conditioned mean. Combined with C1, this means the temporal signal exists but is entangled with length in the monolithic statistic — exactly the pattern the early/late decomposition resolves.

#### 4.3 C3 — Orthogonal-pair-only $\kappa\_t$ (marginal; directionally consistent)

**Hypothesis under test:** all 5 probes are linear classifiers on the *same* L43 residual.  $\kappa\_t$  might therefore be measuring “is this residual stream cleanly linearly separable by *any* axis?” — a one-dimensional quantity of “residual quality” — rather than genuine multi-axial coherence.

**Test:** Greedy selection of probes with maximal mutual decorrelation ( $|\text{corr}| < 0.20$  to all already-selected probes); recompute  $\kappa\_t$  restricted to this subset.

**Result:** 4 of 5 probes selected as orthogonal (A\_tool\_finish, C\_long\_thinking, E\_repo\_ansible, B\_tool\_bash; D was excluded due to its  $-0.83$  correlation with B).

- Monolithic slope, success (orth-only):  $+0.0027$
- Monolithic slope, failure (orth-only):  $+0.0008$
- Effect ratio:  $3.2\times$
- Mann-Whitney  $p$ :  $0.24$

**Verdict: directionally consistent (success > failure with same ratio sign and magnitude), but not significant at the orth-only subset.** The encoder-shared-artifact critique is partially addressed (slope still goes in the same direction with the most-decorrelated subset) but is not closed. We retain it as a limitation.

#### 4.4 C4 — Class trace-length distribution (information)

Documented as the reason C1 and C5 exist.

- Success traces (N=40): median `n_turns` = 37, mean = 35.5
- Failure traces (N=59): median `n_turns` = 50, mean = 50.0
- Mann-Whitney *p* on length:  $2.7 \times 10^{-21}$  (extremely different)
- All 59 failures terminate via `max_turns` exhaustion; all 40 successes terminate via `finish_tool`

This length asymmetry is precisely what makes C1 the critical control. Without the early/late half decomposition, the headline statistic is dominated by the length effect.

#### 4.5 C5 — Iso-length stratum (insufficient power)

**Hypothesis:** within the subset of traces that all hit `max_turns` (a length-constant stratum), does the slope still differ by class?

**Result:** the success/failure split among `max-n_turns` traces is severely imbalanced (1 success, 0 failures at the maximum-length tier under the available stratification window), and no test can be run.

We do not draw conclusions from C5. The early/late half decomposition (§3.2) is the substantively cleaner control because it normalizes length per-trace rather than restricting to a constant-length stratum.

#### 4.6 Summary

Control	Verdict	What it told us
C1 length-regression	<b>Fail</b>	Monolithic slope confounded with length; motivates U-shape decomposition
C2 turn-order shuffle	<b>Pass</b> ( $p < 0.0001$ )	Temporal ordering matters; signal is genuinely sequential
C3 orthogonal-pair	Marginal (directional, $p = 0.24$ )	Encoder-shared-artifact critique partially addressed
C4 length distribution	Information ( $p = 2.7e-21$ )	Failure traces 40% longer than success
C5 iso-length stratum	Insufficient power	Cannot run within max-turns stratum

**Outcome:** the pre-registered original four gates (G1-G4) deliver a mean- $\bar{\kappa}$  discrimination finding ( $\bar{\kappa}$  AUROC 0.677,  $p = 0.0009$ ). The pre-registered five controls (C1-C5) reveal the bonus monolithic-slope statistic is length-confounded, and motivate the length-normalized early/late half decomposition reported in §3.2 as the substantive temporal finding. The temporal claim — that  $\kappa_t$  shows a U-shape whose amplitude differs by class — is therefore post-hoc relative to the original v2 pre-registration, but rests on a control battery whose pass/fail decisions were pre-registered.

## 5. Discussion

### 5.1 Explore-then-consolidate

The U-shape pattern admits a natural mechanistic interpretation. In the early turns of a trace, the agent is in an *exploration* mode: it probes the environment (reads, lists, searches), tries multiple framings of the problem, and generally produces probe outputs that are uncoupled — different turns invoke different tools,

reasoning length varies as the model explores, and tool successes are sporadic. Probe scores therefore *de-correlate* from any initial baseline;  $\kappa_t$  *falls*.

As the trace progresses, a successful agent eventually identifies the change to make and the action sequence to apply it. Once committed to this plan, behavior becomes *consolidated*: tool calls follow a coherent sequence (read  $\square$  edit  $\square$  test  $\square$  finish), reasoning length tracks the complexity of each step, tool successes become reliable. The probes — which separately track tool class, reasoning length, tool success, and repo identity — begin to *re-couple* because the underlying behavior has become coupled.  $\kappa_t$  *rises*.

In a failed trace, neither phase happens with substantial amplitude. The model thrashes: it doesn't commit fully to an exploration strategy in the early turns, and it never consolidates onto a successful plan in the later turns. The probes remain weakly coupled throughout.  $\kappa_t$  stays approximately flat.

The two phases together — the magnitude of the fall, the magnitude of the rise — are what distinguish successful from failed trajectories. The original “monotonic buildup” framing of v1/v2 conflated the two phases into a single average slope, which the trace-length confound then partially explained away. The U-shape decomposition both fixes the confound and reveals a more coherent mechanistic story.

## 5.2 Inverse-cardiac comparison, revisited

In the cardiac analogue (Morris, 2006+), healthy patients have a *coupled* baseline and decompensation is a *collapse* of that coupling. In our LLM-agent setting, the healthy/successful baseline is itself *oscillatory* — coupling normally falls during exploration and rises during consolidation — and failure is the *absence* of that oscillation, not a collapse from it. The structural insight (cross-channel coupling carries health information) transfers; the dynamical signature differs.

The directional inversion is itself an interpretability lesson: when importing structural insights from biology (ICU vital signs, immune feedback, hormonal regulation) into LLM monitoring, the *quantity to measure* often transfers cleanly while the *predictive directionality* depends on the failure mode of the specific system. Our 2-step lesson: (1) the joint covariance of multiple probes is informative; (2) but what *successful* coupling looks like — high-static, low-static, oscillatory — is system-specific and requires direct measurement.

## 5.3 Implications for agent monitoring

The U-shape decomposition admits an online implementation. At each new agent turn  $t$ :

1. Run all  $N$  probes on the current turn's residuals.
2. Append the new scores to a sliding window of size  $2W+1$ .
3. Recompute the windowed mean-absolute correlation. Cost: one  $N \times N$  correlation matrix per turn.
4. Maintain *two* exponentially-smoothed estimates: the most-recent slope over the first half of the running trajectory, and the most-recent slope over the second half.

For agent-probe-guard SDK v0.2, we plan to expose  $\kappa_t$ , its early-half slope estimate, and its late-half slope estimate as a three-component meta-signal alongside individual probe outputs. Operational uses:

- **Early-half slope substantially less negative than the per-task baseline**  $\square$  flag for human review (insufficient exploration).
- **Late-half slope substantially less positive than the per-task baseline**  $\square$  flag for escalation (consolidation failing).
- **Both halves flat**  $\square$  highest-priority warning: trajectory is on a likely-failure path with no observable engagement of either phase.

$\kappa_t$  is a *detect* signal, not a *lever*. Whether intervening on  $\kappa_t$  actually changes outcomes (via steering, prompt modification, sampling temperature) is an open question; we do not address it here.



## 5.4 Why the walk-back narrative belongs in the paper

It is unusual to dedicate space in a research paper to documenting one’s own controls catching one’s own headline confound. We do so deliberately:

- The mean- $\bar{\kappa}$  finding (§3.1) was pre-registered and survives all controls. It is the cleanest claim in the paper.
- The U-shape finding (§3.2) is post-hoc relative to the original v2 pre-registration, but the controls battery that *motivated* its construction (C1) was pre-registered.
- The reader should be able to distinguish these two epistemic statuses cleanly. Hiding the C1 fail and presenting only the U-shape finding would falsely advertise the temporal claim as primary; presenting only the mean discrimination would understate the substantive interest.

In our view this kind of two-tier reporting — pre-registered finding + transparent post-hoc rescue — is the right epistemic stance for early-stage interpretability work, where the “ablation that should have worked” failing is the norm, not the exception.

## 5.5 Limitations

1. **Single model.** All findings are on Qwen3.6-27B. Multi-model validation is the most important next step.
2. **Single benchmark.** SWE-bench Pro is a hard agentic coding benchmark; behavior on different agent tasks (web navigation, retrieval-augmented QA, dialog) is untested.
3. **Single layer, single token-position.** Residuals are captured at L43 `paired_concat`. Layer/position sensitivity is an open empirical question.
4. **N=99 traces.** The early-half/late-half p-values ( $10^{-4}$  to  $10^{-5}$ ) are robust but tighter confidence intervals require scale-up. An attempted N=200 scale-up failed at the data-pipeline level (SWE-bench Pro `base_commit` retrieval; documented in our project notes).
5. **C3 encoder-shared-artifact critique partially open.** The orthogonal-pair subset shows the slope effect with the same direction and similar magnitude (3.2× ratio) but not significance ( $p=0.24$ , possibly underpowered with only 4 axes).
6. **U-shape decomposition is post-hoc** relative to the original v2 pre-registration. A pre-registered replication of the early-half/late-half findings on an independent dataset would substantially strengthen the temporal claim.
7. **G4 absolute-value EARLY  $\kappa_t$  fails ( $0.572 < 0.60$ ),** consistent with the U-shape interpretation that early discrimination is in the rate of change, not the level.

---

## 6. Related Work

### 6.1 Multi-probe ensembles for LLM monitoring

Anthropic’s “Cheap Monitors” (Apr 2026) and Persona Vectors develop several variants of a single behavioral axis. Apollo Watcher (Feb–May 2026) focuses on a single LLM-judge classifier per failure mode. OpenAI’s internally-reported monitoring infrastructure (Mar 2026) reports 99.9% coverage from single classifiers.

Hua et al. (2025, arXiv:2507.15886) study combining a probe with a black-box monitor for a single failure mode — useful, but the two signals measure the *same* axis. To our knowledge, no published work computes cross-axis correlation as the meta-signal we report, and no published work computes its early/late half decomposition over agent trajectories.

### 6.2 Cardiac uncoupling and Early Warning Scores

Morris (2006) established the link between cross-vital-sign correlation breakdown and ICU decompensation. The subsequent 20-year literature on Early Warning Scores (MEWS, NEWS, NEWS2) operationalizes

the insight at hospital scale. We borrow the structural insight (cross-channel coupling carries health information) and adapt it; the *dynamical signature* differs in the LLM-agent setting (U-shape vs. coupled baseline), as discussed in §5.2.

### 6.3 LLM reasoning quality probes

Boppana et al. (“Reasoning Theater,” arXiv:2603.05488) report 87% AUROC on a single-phase reasoning-quality probe in a synthetic benchmark. Chen et al. (2025, arXiv:2505.05410) document inner-outer divergence between thinking and answer tokens qualitatively, without explicit probing. Young et al. (arXiv:2603.26410) probe thinking-vs-answer divergence at AUROC 0.554. None compute a cross-probe meta-signal or a temporal decomposition.

### 6.4 Explore-then-exploit in agent reasoning

The exploration–consolidation pattern we identify mechanistically in §5.1 has antecedents in reinforcement-learning theory (the explore-exploit trade-off; Sutton & Barto 2018) and in cognitive-science accounts of human problem-solving (insight-then-search; Ohlsson 1992). To our knowledge, this is the first observation that the same pattern is visible in the residual-stream activations of an LLM agent via a multi-probe meta-signal.

### 6.5 Our own prior and adjacent work

The agent-probe-guard v0.1 SDK (Vicentino, 2026a) ships a 2-probe ensemble (thinking-stage + capability) on Qwen3.6-27B. The Paper-5 saturation-direction work (Vicentino, 2026c) characterizes the *causality* of single probes under steering. The Two-Forms paper (Vicentino, 2026d) documents distinct epiphenomenal regimes single probes occupy. The present work extends the line by moving from single-probe to multi-probe analysis, and by introducing a *temporal* statistic.

---

## 7. Conclusion

We propose cross-probe coherence  $\kappa_t$  as a meta-signal for LLM agent monitoring and characterize two findings on 99 SWE-bench Pro trajectories from Qwen3.6-27B. (1) Per-trace mean  $\bar{\kappa}$  separates outcomes at AUROC 0.677 (pre-registered,  $p=0.0009$ ). (2)  $\kappa_t$  exhibits a U-shape over each trajectory; the *amplitude* of the U distinguishes successful from failed agents in both halves ( $p=0.0002$  early,  $p=0.00004$  late). The U-shape is mechanistically interpretable as an explore-then-consolidate arc: successful agents commit to exploration in the early phase and to consolidation in the late phase; failed agents do neither.

The pattern is *inverse* to its biological inspiration. Cardiac decompensation is a *collapse* from a coupled baseline; LLM-agent failure is the *absence* of an oscillatory baseline. The structural insights from biology transfer cleanly even when the dynamical signature differs.

Our pre-registered controls caught a substantial length-confound in the original monolithic-slope statistic of  $v_2$ ; the U-shape decomposition is the post-control rescue. We document this walk-back-and-rescue trajectory explicitly because we believe pre-registration discipline — not the headline statistic in isolation — is what carries the credibility of an early-stage interpretability result.

The finding is single-model, single-benchmark, single-layer,  $N=99$ . We document these limitations honestly along with the methodological discipline that we believe makes the result worth reporting.

---

## Appendix A. Reproducibility

Code (Apache-2.0):

- `openinterp-swebench-harness/scripts/run_kappa_t_v2.py` (v2 with 5 axes; mean- $\bar{\kappa}$  and monolithic-slope)
- `openinterp-swebench-harness/scripts/run_kappa_t_v3.py` (v3 internal replication, 9 axes)
- `openinterp-swebench-harness/scripts/run_kappa_t_controls.py` (5 robustness controls: C1-C5)
- `openinterp-swebench-harness/scripts/run_kappa_t_c6_length_controlled.py` (early/late half decomposition + length stratification)
- `openinterp-swebench-harness/scripts/make_figures_kappa_t.py` and `make_fig7_controls.py` (figure generation)

#### Data:

- Phase 6 capture set (99 traces, 11 layers, 4 positions, ~9 MB/trace): Hugging Face dataset `openinterp/swebench-pro-phase6-residuals` (forthcoming)
- $\kappa_t$  v2, v3, controls, and c6 result JSONs: HF dataset `openinterp/kappa-t-coherence-buildup`

#### Compute:

- Phase 6 capture: single A100, ~12 hours (already amortized)
- $\kappa_t$  analysis (v2, v3, controls, c6): MacBook Pro CPU, ~35 minutes total
- Total marginal compute for this paper after Phase 6: **R\$0** (no GPU required)

## Appendix B. Walk-back history

Five pre-registered single-probe candidates failed gates in the 36 hours preceding the  $\kappa_t$  v2 work:

1. **Silent-refusal Phase 2A** (2026-05-17): irreproducible across `transformers-commit` drift (Phase 1 vs Phase 2A used different commits from `main` 17h apart;  $P=3.4e-6$  for the  $12 \times 0$  silent-count change under chance).
2. **Tool-doubt Phase 1** (2026-05-17): 40 (layer, position) probe sweep; 0 combinations with AUROC gap > 0.15. Top-5 inspected residuals dominated by `Error: / ok=false` substrings  $\square$  lexical leak.
3. **Tool-doubt Phase 1b position sweep** (2026-05-17): diagnostic confirms Qwen3.6-27B residual encodes explicit errors lexically (L7–L11) but is blind to silent corruption at all layers/positions.
4. **Inner-Outer paired probe** (2026-05-18): paired-residual probe at `think_end` + `turn_end`. AUROC barely exceeded single-residual probe (+0.023 lift); top-10 inspected residuals dominated by `max_turns` traces  $\square$  capability redux.
5.  **$\kappa_t$  v1 with trace-level labels** (2026-05-18 morning): probes given trace-constant labels collapsed within-trace variability; G4 EARLY  $\kappa_t = 0.458$ .

The  $\kappa_t$  v2 finding (§3.1 of this paper) emerged from re-labeling probes with per-turn rather than per-trace labels. The U-shape finding (§3.2) emerged in turn from the C1 control catching a length-confound in v2’s bonus slope statistic. The pattern of progressive walk-back-and-rescue is what we treat as load-bearing for credibility.

## Appendix C. Software versions

- Python 3.9 (macOS) for analysis; Python 3.11 (Colab) for Phase 6 capture
- `transformers` from `main`, commit `73d9159697a851c85623d0f03fcfbdd863d38975` (Phase 6 capture); reanalysis is residual-only and version-independent
- `numpy` 1.26, `scikit-learn` 1.4, `scipy` 1.13, `safetensors` 0.4
- GPU: NVIDIA A100 80GB (Colab Pro+) for Phase 6 capture; CPU only for analysis

## Appendix D. Figures

See `paper/figures/`:

- `fig1_kappa_t_trajectories.png` — per-class mean  $\kappa_t$  trajectories with bootstrapped 95% bands

- `fig2_slope_distributions.png` — per-trace monolithic-slope distributions, success vs failure (now superseded by fig7’s early/late panels)
  - `fig3_probe_aurocs.png` — individual probe AUROCs (v2 and v3 sets)
  - `fig4_probe_correlations.png` — probe-score pairwise correlation heatmap
  - `fig5_mean_kappa_violin.png` — per-trace mean  $\bar{\kappa}$  violin plot by class
  - `fig6_v2_vs_v3_replication.png` — v2 vs v3 gate-by-gate comparison
  - (forthcoming) `fig7_controls_summary.png` — robustness controls C1–C5 panel
  - (forthcoming) `fig8_early_late_half.png` — early-half/late-half slope distributions by class (primary new figure)
- 

## References

- Anthropic. *Cheap Monitors*. Apr 2026. <https://www.anthropic.com/research/cheap-monitors>
- Anthropic. *Persona Vectors*. 2026. <https://www.anthropic.com/research/persona-vectors>
- Apollo Research. *Watcher series*. 2026.
- Boppana, A. et al. *Reasoning Theater*. arXiv:2603.05488, 2026.
- Chen, J. et al. *Inner-outer divergence in chain-of-thought reasoning*. arXiv:2505.05410, 2025.
- Hua, X. et al. *Optimally combining probe and black-box monitors*. arXiv:2507.15886, 2025.
- Morris, A. *Cross-vital-sign decorrelation in ICU decompensation*. Crit. Care Med., 2006.
- Ohlsson, S. *Information-processing explanations of insight*. In Keane & Gilhooly (eds.), *Advances in the Psychology of Thinking*, 1992.
- OpenAI. *Internal monitoring infrastructure*. OpenAI blog, Mar 2026.
- Sutton, R. & Barto, A. *Reinforcement Learning: An Introduction*. 2nd ed., MIT Press, 2018.
- Vicentino, C. (2026a). *agent-probe-guard v0.1: detection-only multi-probe SDK for LLM agents*. [openinterp.org/research/papers/agent-probe-guard](https://openinterp.org/research/papers/agent-probe-guard).
- Vicentino, C. (2026b). *SWE-bench Pro failure anatomy on Qwen3.6-27B: Phase 6 capture protocol*. [github.com/OpenInterpretability/openinterp-swebench-harness](https://github.com/OpenInterpretability/openinterp-swebench-harness).
- Vicentino, C. (2026c). *Saturation-direction probe levers*. [openinterp.org/research/papers/saturation-direction-probe-levers](https://openinterp.org/research/papers/saturation-direction-probe-levers).
- Vicentino, C. (2026d). *Two forms of epiphenomenal probes*. [openinterp.org/research/papers/two-forms-epiphenomenal-probes](https://openinterp.org/research/papers/two-forms-epiphenomenal-probes).
- Young, K. et al. *Thinking-answer divergence probing*. arXiv:2603.26410, 2026.