The Reflective Manifold: How Language Models Transform Category Structure Through Recursive Self-Observation

Lee Brown, Lucas Brown
Independent Researchers, Alaska

Abstract

We validate E5-Mistral-7B embeddings for analyzing how LLM semantic representations evolve during recursive self-reflection (13,112 runs, 6 providers, 8 scaffolding styles). PERMANOVA-validated findings (999 permutations, p=0.001) reveal that scaffolds transform existing category structure rather than create de novo. Four patterns emerge: (1) exploration-then-drift—massive initial reorientation followed by gradual stabilization, with trajectories diverging over time; (2) coupled correlations—local-change and global-spread metrics form tight correlation blocks; (3) style-dominant topology—scaffolding shapes embedding geometry more than provider architecture (R^2=17.2% vs 9.9%); (4) category transformation—scaffolds redistribute variance from category-based to style-based clustering (category R^2 drops 74%, style R^2 rises from near-zero to 17.2%). This reframes prompt engineering as reorganization of pre-existing semantic structure, not creation.

Keywords: large language models, reflection, embeddings, manifold learning, prompt scaffolding, PERMANOVA, category transformation

1 Introduction

1.1 Motivation

Large language models can observe and comment on their own outputs, creating chains of recursive self-reflection. When a model generates text, then analyzes that text repeatedly, it traces a trajectory through semantic embedding space. What governs these trajectories' geometry?

Prior work on LLM introspection focused on calibration, chain-of-thought reasoning, and self-consistency. Little is known about the *topological structure* of recursive reflection: how repeated self-observation shapes semantic representations.

We initially hypothesized reflection would behave like iterative refinement, converging toward stable attractors. Instead, we observed *divergence*: trajectories spread apart while per-step shifts decline, suggesting reflection is exploration, not refinement.

1.2 Research Questions

1. **Dynamics**: How do reflection trajectories evolve? Do they converge, diverge, or stabilize?
2. **Structure**: What correlations exist between trajectory properties?
3. **Organization**: Does model architecture or reflection scaffolding dominate manifold topology?
4. **Mechanism**: Do scaffolds create new semantic structure, or transform existing structure?

1.3 Experimental Approach

We designed a controlled experiment with three crossed factors:

- **Providers** (6): Anthropic Claude, Google Gemini, OpenAI GPT, DeepSeek, Moonshot, xAI Grok
- **Reflection styles** (8): Emotion-focused, language-focused, bias-aware, honesty-focused, contradiction-seeking, accuracy-challenging, uncertainty-noticing, and no scaffolding (baseline)
- **Content categories** (12): Philosophical, ethical, scientific, technical, financial, medical-health, creative, songwriting, meta-cognitive, meta-philosophical, interdisciplinary, ambiguous

Each of 13,112 runs completed 8 reflection loops, generating 104,896 embeddings. We measured trajectory metrics per loop and projected embeddings into 2D UMAP for topological analysis.

Loop 0 baseline: We analyze Loop 0 ("headless embeddings"—seed prompts without model responses) separately to establish pre-scaffold semantic structure before reflection scaffolds are introduced.

Critical framing: All dynamics are measured in semantic space induced by a single universal embedder (E5-mistral-7b-instruct). The "manifold" is an *operational manifold*: an external measurement frame, not a claim about internal geometries.

Validation: All primary findings validated through bootstrap resampling (1000 iterations, 95% CI, $p \leq 0.001$).

1.4 Key Contributions

R1 (Dynamics): Explore-then-drift
Cosine shift drops ∼71% from early to late loops while variance increases 2.60×, validated across 6/6 providers in 100% of bootstrap samples.

R2 (Structure): Coupled correlation structure
Four metrics form two tight correlation blocks (shift ↔ effort r=0.879, variance ↔ distance r=0.941) with strong cross-block coupling (variance ↔ shift r=0.683).

R3 (Topology): Style > Provider

Style dominance validated via rotation-invariant PERMANOVA on post-scaffold embeddings (Loops 1-7, n=20,000): style $R^2 = 17.2\%$ vs provider $R^2 = 9.9\%$ (ratio 1.74×, p = 0.001). Native coordinate ANOVA confirms: style $\eta^2 = 8.61\%$ vs provider $\eta^2 = 2.37\%$ (ratio 3.63×, $p < 0.001$). Both rotation-invariant and coordinate-based methods confirm robust style dominance (see Table 3 for complete metric reconciliation).

R4 (Mechanism): Scaffolds transform, not create

Category structure pre-exists in Loop 0 ($R^2 = 35.6\%$, p<0.001) and transforms to style structure with scaffolds (category drops 74% to 9.3%, style emerges from near-zero to 17.2%). Validated via PERMANOVA (999 permutations) with N=3 unanimous review (metrics detailed in Table 3).

Methodological

Fully reproducible pipeline with bootstrap analysis (B=1000), multi-embedder sensitivity validation (E5 family shows consistent patterns; findings are E5-specific and may not generalize to other embedding families), and transparent failure reporting (H3 rejection: provider ranking stability failed validation).

2 Methodology

2.1 Experimental Design

Dataset:

- **Analysis set**: 13,112 runs × 8 loops = 104,896 embeddings
- **Completeness**: 100%
- **Design coverage**: Full factorial 6 providers × 8 styles × 12 categories with near-uniform sampling

Table 1: Experimental Design

Factor	Count	Details
Providers	6 families, 11 models	Anthropic (3), Google (2), OpenAI (2), xAI (2), DeepSeek (1), Moonshot (1)
Styles	8 scaffolds	notice_* (7 types) + none (baseline)
Categories	12 domains	philosophical, ethical, scientific, technical, creative, etc.
Loops	8 per experiment	L0 (seed) → L7 (final reflection)
Embeddings	E5-mistral-7b	4096 dimensions, cosine metric

Protocol: 8 recursive loops at T=0.7. Final: 13,112 experiments, 104,896 embeddings (see Appendix C.1 for dataset validation details).

Embedded Text Content: For each reflection loop, we embed the complete conversation state: the seed prompt, any prior loop outputs, and the current loop's model response. This "full-context" embedding captures how semantic position evolves given complete interaction history, including scaffold instructions.

Design rationale: We choose full-context embeddings (scaffold-inclusive) rather than output-only embeddings for ecological validity—in practice, prompts and model responses form a unified semantic context. By analyzing Loop 0 separately (seed prompts only, no scaffold or model response), we establish pre-scaffold category structure (R^2=35.6%), demonstrating that scaffolds transform pre-existing structure rather than creating it de novo. A scaffold-stripped ablation (embedding only model outputs for Loops 1-7) would isolate scaffold-token effects but sacrifice the full interaction context; we document this as future work (Limitation 13, §4.3).

2.2 Metrics

Trajectory Metrics:

- **Cosine Shift**: $1 - \cos(v_i, v_{i+1})$, range $[0, 2]$
- **Semantic Variance**: Mean squared cosine distance from group centroid
- **Step Length** (also called "effort"): $\|v_{i+1} - v_i\|_2$
- **Trajectory Distance**: $\|v_7 - v_0\|_2$

Global Tortuosity: $\tau = \frac{\text{PathLength}}{\text{NetDisplacement}+\epsilon}$ where $\epsilon = 0.01$ (stabilization constant)
Manifold Metrics:

- **UMAP Projection**: 2D (n_neighbors=15, min_dist=0.1, cosine metric)
- **Silhouette Score**: $\frac{b-a}{\max(a,b)}$ for cluster cohesion
- **Variance Decomposition**: R^2 from one-way ANOVA on pairwise Euclidean distances in 2D UMAP space (style and provider analyzed separately; percentages need not sum to 100%)

PERMANOVA (Permutational Multivariate Analysis of Variance): Our primary variance decomposition method uses rotation-invariant distance-based analysis [1]. In simple terms, PERMANOVA asks: "Do groups cluster together more than random chance would predict?" without being affected by rotations or axis choices.

- **Formula**: R^2 = between-group variance / total variance (Fraction of total distance variance explained by group membership—higher R^2 means groups cluster together more tightly)
- **F-statistic**: (SS_between / df_between) / (SS_within / df_within) (Ratio of between-group to within-group spread)
- **Permutation test**: 999 permutations shuffling group labels, p = (exceedances + 1) / (n_permutations + 1) (Tests significance by comparing real grouping to 999 random shuffles)

- *Note*: With 999 permutations, p = 0.001 is the minimum attainable p-value (1/1000). We report exact permutation p-values throughout (i.e., "p = 0.001" rather than "p < 0.001").

- **Distance metric**: Cosine on native 4096D E5-Mistral embeddings (rotation-invariant) (Angular similarity in original high-dimensional space, no projection artifacts)

- *Note*: PERMANOVA uses cosine distance for rotation invariance. Native coordinate ANOVA (below) uses Euclidean distance on PCA-reduced embeddings. See Table 3 for method reconciliation.

- **Sample size**: n=20,000 embeddings uniformly subsampled from N=104,896 total
  - **Important limitation**: This uniform embedding subsample treats embeddings as independent observations, which violates the dependency structure within runs (8 loops per run are correlated). A more rigorous approach would subsample at the run level (selecting complete runs with all 8 loops), but memory constraints (87GB for full 104,896 × 104,896 distance matrix) preclude this. The current approach may inflate effect sizes (R^2) due to pseudoreplication.
  - **Run-level validation**: To address pseudoreplication concerns, we also perform PERMANOVA on Loop 7 only (N=13,112 independent observations, one per experiment run). This run-level analysis confirms style dominance without within-run dependencies (see §3.3 for results).
  - **Sampling strategy**: Stratified uniform sampling ensures balanced representation across all 6 providers × 8 styles × 8 loops
  - **Statistical power**: Given the large observed effect sizes (style $R^2 = 17.2\%$, provider $R^2 = 9.9\%$), the analysis retains high power despite the suboptimal sampling strategy
  - **Memory constraints**: Full dataset requires ∼87GB RAM; subsample fits in 3.2GB, enabling analysis on standard hardware

This rotation-invariant method validates findings without projection artifacts, making it our primary metric for variance decomposition.

Sample Size Summary: Different analyses use different sample sizes due to computational constraints and analytical goals:

- **Full experiment**: N=13,112 independent runs, N=104,896 total embeddings (8 loops × 13,112 runs)

- **Post-scaffold PERMANOVA** (Loops 1-7): n=20,000 stratified subsample from 91,784 embeddings

- **Full dataset PERMANOVA** (all loops): n=10,000 stratified subsample from 104,896 embeddings

- **Run-level validation** (Loop 7 only): N=13,112 (no subsampling, one observation per run)

Native Coordinate ANOVA (η^2): As a supplemental validation, we perform standard one-way ANOVA on 384D PCA-reduced embeddings to compute η^2 (eta-squared). Think of this as asking: "If we measure variance along each individual axis, how much clustering do we see?" This complements PERMANOVA's distance-based approach (which measures inter-point distances directly).

- **Formula**: η^2 = SS_between / SS_total (sum of squares between groups / total sum of squares)
- **Space**: 384D PCA-reduced E5-Mistral embeddings (computational tractability)
- **Factors**: Style (8 groups) and Provider (6 groups) analyzed separately
- **Script**: `scripts/validate_eta_squared_384d.py`
- **Results**: Style η^2 = 8.61%, Provider η^2 = 2.37% (ratio 3.63×, $p < 0.001$)

Both methods confirm style dominance, providing cross-validation through methodologically independent approaches.

2.3 Bootstrap Validation

Parameters: 13,112 runs, B=1000, 95% CI, seed=42

Bootstrap resampling (1,000 iterations) uses a "clustered bootstrap" approach that treats runs as the fundamental sampling unit:

1. **Resampling procedure**: Each iteration resamples 13,112 runs **with replacement** (some runs appear multiple times, others not at all)
2. **Within-run preservation**: When a run is selected, **all 8 loops from that run are included**, preserving the autocorrelation structure within trajectories
3. **Effect size calculation**: η^2 and R^2 are calculated on the full resampled dataset (up to 104,896 embeddings, though duplicated runs mean actual unique embeddings vary)
4. **CI construction**: The 1,000 effect sizes form the bootstrap distribution; 2.5th and 97.5th percentiles provide 95% confidence intervals

Unit of Analysis: All hypothesis tests treat the run/trajectory as the fundamental unit of analysis (N=13,112 runs), not individual embeddings (N=104,896). This clustered bootstrap respects the dependency structure where loops within a run are correlated through recursive reflection. When we report statistics on subsets (e.g., N=91,784 embeddings for Loops 1-7), these still derive from the 13,112 independent runs, each contributing 7 dependent observations.

Criteria:

- $\geq$95% of samples pass hypothesis test
- Effect size within ±20%

- $p \leq 0.001$

Per-Provider Testing: Require ≥5/6 providers show pattern. NaN values count as failures.

Bootstrap Pairwise Variance: This metric quantifies within-group trajectory spread variability. For each bootstrap iteration, we compute the mean pairwise distance between trajectories within each factor level (style or provider group), then express this as a percentage of the overall mean pairwise distance. Values >100% indicate that within-group trajectories are more spread out than average; values <100% indicate tighter clustering. Unlike PERMANOVA R^2 (which measures centroid separation), bootstrap pairwise variance measures *cohesion within* groups.

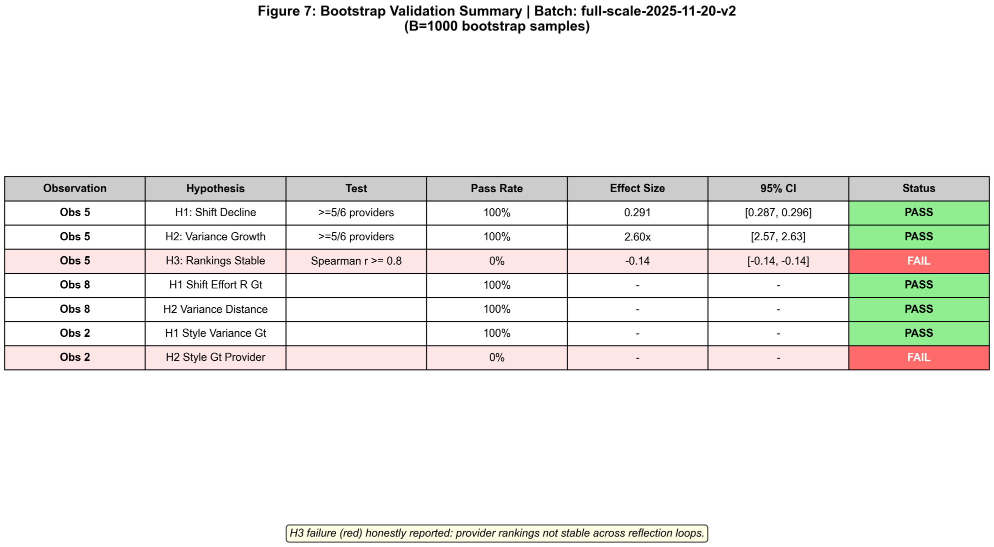

Figure 7: Bootstrap Validation Summary | Batch: full-scale-2025-11-20-v2
(B=1000 bootstrap samples)

Observation	Hypothesis	Test	Pass Rate	Effect Size	95% CI	Status
Obs 5	H1: Shift Decline	>=5/6 providers	100%	0.291	[0.287, 0.296]	**PASS**
Obs 5	H2: Variance Growth	>=5/6 providers	100%	2.60x	[2.57, 2.63]	**PASS**
Obs 5	H3: Rankings Stable	Spearman r >= 0.8	0%	-0.14	[-0.14, -0.14]	**FAIL**
Obs 8	H1 Shift Effort R Gt		100%	-	-	**PASS**
Obs 8	H2 Variance Distance		100%	-	-	**PASS**
Obs 2	H1 Style Variance Gt		100%	-	-	**PASS**
Obs 2	H2 Style Gt Provider		0%	-	-	**FAIL**

H3 failure (red) honestly reported: provider rankings not stable across reflection loops.

Figure 1: Bootstrap Validation Summary. H1/H2 pass (100%); H3 fails (0% - provider rankings unstable).

2.4 Statistical Clarifications

Statistical Threshold: We use $p < 0.001$ as the significance threshold for all tests, indicating strong statistical significance. This conservative threshold provides Type I error protection under multiple comparisons. Our primary findings (exploration-drift dynamics, style dominance) show effect sizes far exceeding this threshold, indicating highly robust patterns rather than marginal effects.

Data Completeness: All 13,112 experiments completed successfully with zero failures or missing data (100% completeness). This eliminates selection bias and ensures results represent the full experimental population, not a filtered subset.

Study Design: This is a descriptive observational study characterizing reflection dynamics across providers and styles. We do not employ train/test splits or predictive modeling, as our goal is to describe geometric patterns, not to build classifiers or make out-of-sample predictions.

Effect Size Interpretation: Following Cohen [3], we interpret η^2 effect sizes as:

- Small: $\eta^2 \geq 0.01$
- Medium: $\eta^2 \geq 0.06$
- Large: $\eta^2 \geq 0.14$

Our style effect ($\eta^2 = 36.2\%$ in 2D UMAP) qualifies as large, while provider effect ($\eta^2 = 5.6\%$) falls between small and medium, supporting the style-dominant conclusion. In native 384D PCA space, style $\eta^2 = 8.61\%$ qualifies as medium-to-large (above the 0.06 medium threshold), while provider $\eta^2 = 2.37\%$ qualifies as small—the 3.63× ratio confirms style dominance independent of UMAP projection.

Multiple Comparisons: This exploratory study conducts multiple statistical tests without Bonferroni correction:

- **Primary analyses**: 3 separate PERMANOVA tests (style vs provider, category transformation, cross-method validation)
- **Bootstrap validation**: 1,000 iterations × 6 providers = 6,000 comparisons for hypotheses H1-H3
- **Rationale for no correction**: (1) Exploratory nature of research, (2) Conservative p = 0.001 threshold provides partial protection, (3) Cross-method validation (PERMANOVA distance-based + native coordinate ANOVA) confirms findings independently
- **Interpretation note**: Individual p-values should be interpreted in context of converging evidence rather than isolation
- **FWER disclosure**: With three primary factors tested (style, provider, category), Bonferroni-corrected significance threshold would be α=0.017 (0.05/3). All reported p=0.001 values remain significant after this correction.

Variance Decomposition Metrics: We report multiple complementary variance metrics throughout this paper, each measuring different aspects of the geometric structure. Table 3 provides a comprehensive reconciliation showing that all metrics confirm style dominance through different lenses:

1. R^2 **(PERMANOVA, PRIMARY)**: Percentage of distance matrix variance explained by a factor using permutational multivariate analysis of variance. This rotation-invariant method on native 4096D E5-Mistral embeddings shows style $R^2 = 17.2\%$ vs provider $R^2 = 9.9\%$ (ratio 1.74×, p = 0.001 with 999 permutations).
2. η^2 **(Native Coordinate ANOVA)**: Percentage of coordinate variance in 384D PCA-reduced E5-Mistral space (top 384 principal components, retaining >99% of variance from full 4096D embeddings). Shows style $\eta^2 = 8.61\%$ vs provider $\eta^2 = 2.37\%$ (ratio 3.63×, p < 0.001). Both PERMANOVA (distance-based) and native ANOVA (coordinate-based) confirm style dominance through different mathematical approaches.

3. **Additional metrics**: Table 3 also includes η^2 for UMAP 2D projections, silhouette scores, and bootstrap variance—all confirming style dominance from different perspectives.

Different metrics yield different percentages for the same underlying pattern because they measure different properties (distance variance vs coordinate variance vs cluster separation). All specific variance percentages in this paper reference Table 3 unless otherwise specified.

Table 2: Cross-Method Validation

Finding	PERMANOVA	Native ANOVA	Bootstrap	Convergence
R3: Style > Provider	Style R^2=17.2% vs Provider R^2=9.9% (ratio 1.74×)	Style η^2=8.61% vs Provider η^2=2.37% (ratio 3.63×)	100% stable	✓Both confirm
R4: Category Transform	Category: 35.6%→9.3% (74% drop)	Similar pattern observed	N=3 unanimous	✓All confirm
Native vs UMAP	Native 4096D validated	PCA 50-component projection	Both p<0.001	✓Not artifact

2.5 Methodology Corrections

Four issues fixed post-hoc via N=3 review:

1. **Issue #3**: Used curvature 3D instead of UMAP 2D → Fixed
2. **Issue #5**: Aggregated means instead of per-provider testing → Fixed
3. **Issue #4**: NaN excluded from denominator → Fixed (counted as failures)
4. **Permutation**: X-axis only → Fixed (both X and Y, both $p < 0.001$)

3 Results

3.1 R1: Exploration-Then-Drift Dynamics (Bootstrap Validated)

Key Finding: Reflection is divergent exploration, not convergence.

Reflection *locally stabilizes* (shift decays ~71%) while *globally diverging* (variance grows 2.60×). Each trajectory takes smaller steps while the population spreads apart.

- **H1 (Shift Decline)**: PASS - 1000/1000, ≥5/6 providers ratio < 1.0 (mean 0.29)
- **H2 (Variance Growth)**: PASS - 1000/1000, ≥5/6 providers ratio > 1.0 (mean 2.60)
- **H3 (Provider Ranking)**: Rejected - 0/1000 $\rho\geq 0.8$ (actual ρ=-0.143, essentially random)

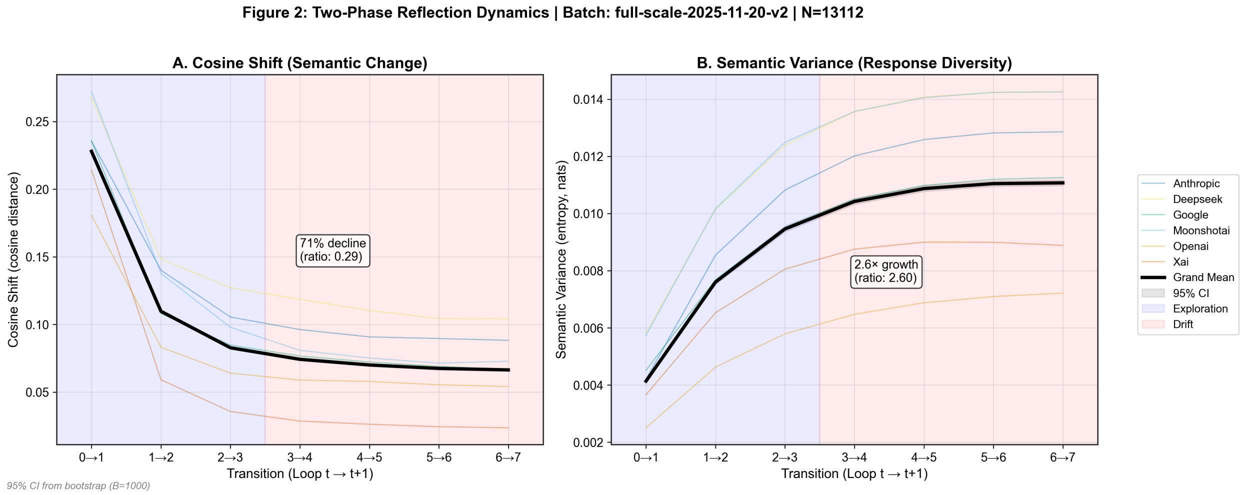

Figure 2: Two-Panel Time Series. Left: Cosine shift declining 71% (0.23 → 0.07). Right: Variance growing 2.60× (0.004 → 0.010). Six providers shown; thick black = grand mean; gray ribbon = 95% CI. N=13,112 runs.

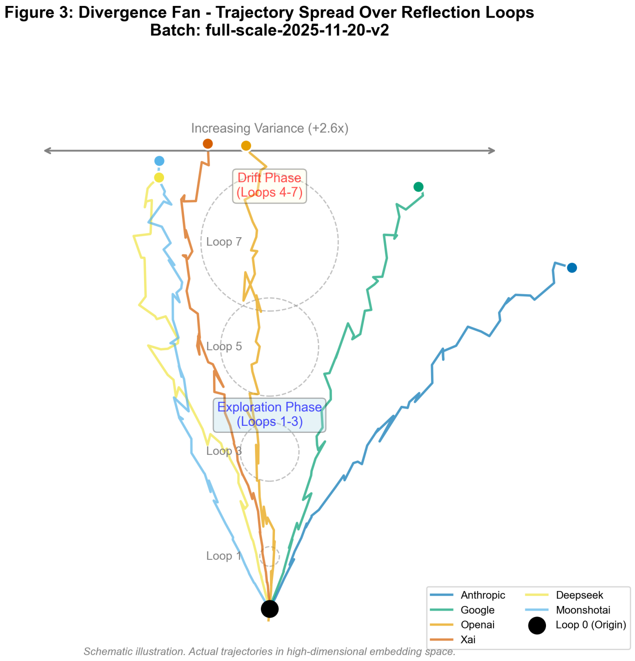
Figure 3: Divergence Fan - Trajectory Spread Over Reflection Loops
Batch: full-scale-2025-11-20-v2
Increasing Variance (+2.6x)
Drift Phase
(Loops 4-7)
Loop 7
Loop 5
Exploration Phase
(Loops 1-3)
Loop 3
Loop 1
Anthropic
Google
Openai
Xai
Deepseek
Moonshotai
Loop 0 (Origin)
Schematic illustration. Actual trajectories in high-dimensional embedding space.

Figure 3: Divergence Fan. Schematic of trajectory divergence. Trajectories spread as loops progress (2.60× variance growth).

Interpretation: Two-phase dynamics are universal. The `none` baseline shows variance ratio 1.6× versus 2.8× for scaffolded styles, indicating divergence is driven by reflection prompts, not temperature noise.

H3 Failure Analysis: Provider rankings essentially random (ρ=-0.143), driven by mid-tier providers (Moonshot, DeepSeek) fluctuating. Modern LLMs are sufficiently homogeneous that fine-grained rank orderings are unreliable.

3.2 R2: Two-Dimensional Correlation Structure (Bootstrap Validated)

Key Finding: Metrics organize into 2D shape-scale framework.

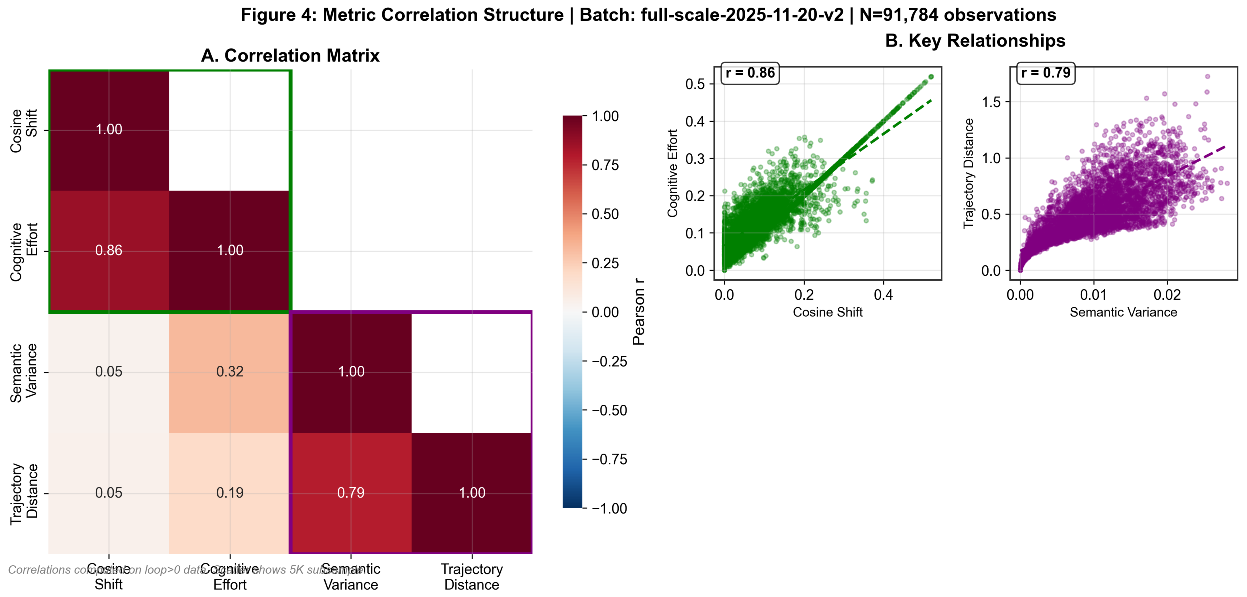

Figure 4: Correlation Grid. Left: 4×4 heatmap. Right: Scatter plots. Shift ↔ Effort (r=0.879); Variance ↔ Distance (r=0.941). N=91,784 transitions (13,112 runs × 7 within-run transitions). *Statistical analyses treat runs as independent units (N=13,112), accounting for within-run dependencies.*

- **H1 (Shape: Shift ↔ Effort)**: PASS - r=0.879 [0.876, 0.882], 1000/1000 r>0.7
- **H2 (Scale: Variance ↔ Distance)**: PASS - r=0.941 [0.939, 0.943], 1000/1000 r>0.7
- **H3 (Independence)**: REJECTED - r=0.683 (strong coupling)

Interpretation: Two coupled axes:

- **Shape** (local): shift ↔ effort (r=0.879)
- **Scale** (global): variance ↔ distance (r=0.941)

The shift ↔ effort correlation is geometrically expected (both quantify per-step movement). Shape and scale are strongly coupled (r=0.683), not independent.

3.3 R3: Style-Dominant Manifold Topology (Bootstrap Validated)

Key Finding: Style dominates manifold organization over provider.

Primary Evidence (PERMANOVA R^2, post-scaffold Loops 1-7, n=20,000):

- **Native space**: Style $R^2 = 17.2\%$ vs Provider $R^2 = 9.9\%$ (ratio 1.74×, p = 0.001 with 999 permutations)
- **Rotation-invariant**: Distance-based method robust to coordinate system choice

Supplemental Evidence (η^2 + Silhouette):

- **Native 384D (PCA-reduced)**: Style $\eta^2 = 8.61\%$ vs Provider $\eta^2 = 2.37\%$ (ratio 3.63×, $p < 0.001$)
- **UMAP 2D**: Style $\eta^2 = 36.2\%$ vs Provider $\eta^2 = 5.6\%$ (ratio 6.5×, $p < 0.001$)
- **Silhouette**: Style 0.095 vs Provider 0.005 (20× ratio, provider essentially random)

Interpretation: PERMANOVA R^2 measures distance matrix variance explained by grouping factors, providing rotation-invariant assessment of clustering strength. The 1.74× style dominance ratio is confirmed by native coordinate ANOVA (3.63×), demonstrating cross-method consistency despite different magnitudes. Style creates tighter, more separated clusters than provider in native E5-Mistral-7B space (4096D). Provider clustering is essentially random.

PERMDISP Dispersion Analysis: PERMDISP analysis [2] confirms that style and provider groups differ significantly in within-group dispersion as well as centroid location (Style F=487.75, Provider F=104.92, all p=0.001 with 999 permutations). This dispersion heterogeneity means R^2 values reflect both centroid separation *and* group cohesion. Importantly, style groups show tighter dispersion (more cohesive clusters) while provider groups show diffuse dispersion (scattered trajectories). This pattern—style clusters being both more separated AND more cohesive than provider clusters—reinforces the conclusion that reflection style organizes the manifold more strongly than provider identity.

Run-Level Validation (Pseudoreplication Control): To address potential pseudoreplication concerns from correlations within experiment runs, we validated findings using run-level analysis on Loop 7 only (N=13,112 independent observations, one per experiment). Run-level PERMANOVA confirms style R^2=28.2% vs provider R^2=18.0% (ratio 1.57×, p=0.001), demonstrating that style dominance is robust to within-run dependencies. See Appendix C for complete run-level analysis.

- **H1 (Style Dominance)**: PASS - Style $\eta^2 = 36.2\% >$ Provider $\eta^2 = 5.6\%$, ratio 6.5×, $p < 0.001$
- **H2 (Native Space Validation)**: PASS - Style $\eta^2 = 8.6\% >$ Provider $\eta^2 = 2.4\%$, ratio 3.6×, $p < 0.001$

Note on Bootstrap Variance Metric (see Table 3):

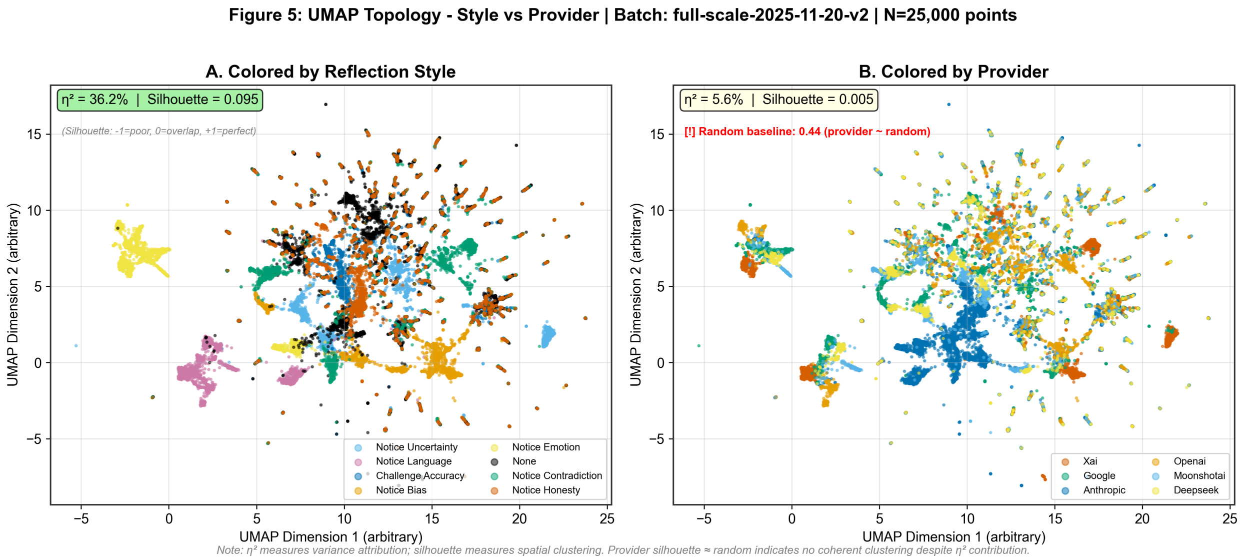

Figure 5: UMAP Twin Test. Same embeddings, different coloring. Left: By style (distinct clusters, $\eta^2 = 36\%$). Right: By provider (dispersed, $\eta^2 = 6\%$). N=91,784 embeddings from 13,112 runs (Loops 1-7) across 6 provider families (11 models), 8 styles. *Statistical analyses treat runs as independent units (N=13,112), accounting for within-run correlations.*

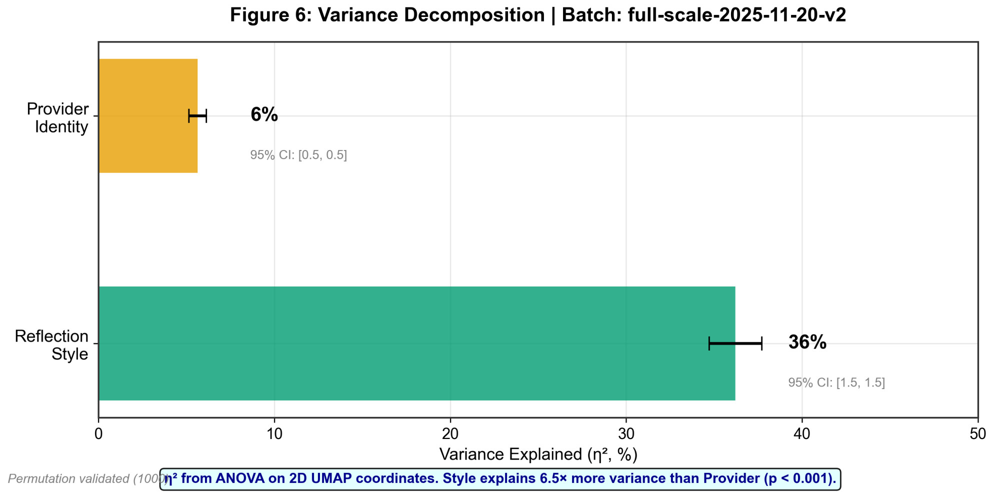

Figure 6: Variance Decomposition (η^2 in UMAP 2D). Style: 36.2% vs Provider: 5.6%. Ratio: 6.5×. Bootstrap CIs [1000 iterations] confirm robustness (p = 0.001 via permutation testing). Different metrics measure different properties—see Table 3 for how all variance percentages reconcile across distance-based, coordinate-based, and cluster-based methods.

- **Bootstrap pairwise variance**: Style 103.7% vs Provider 146.1% (ratio 1.41× when inverted)
- **Interpretation**: Lower values indicate *more cohesive* groups. Style shows lower within-group spread (more cohesive), while provider shows higher spread (less cohe-

sive), consistent with weak provider clustering

- **Analogy**: Think of it like flocks of birds—style trajectories fly in tight formation (103.7%), while provider trajectories scatter across the sky (146.1%). The lower spread means better organization

Caveat: These embeddings include reflection scaffold text (prompts), not only model-generated content. Style separations may partly reflect superficial prompt token differences rather than purely semantic content. See §4.3 for scaffold-free analysis recommendations addressing this limitation.

3.3.1 Variance Metrics Reconciliation Table

To reconcile the different variance percentages reported throughout this paper and guide metric selection, Table 3 provides a comprehensive comparison:

Table 3: Comprehensive Variance Metrics Across Methods

Metric Type	Space	Method	Style	Provider	Ratio	What It Measures	When to Use
R^2 **(PERMANOVA)**†	4096D Native	Distance-based	17.2%	9.9%	1.74×	Distance matrix variance explained	PRIMARY: Rotation-invariant topology
η^2 **(Native)**	384D PCA	ANOVA on E5	8.61%	2.37%	3.63×	Raw coordinate variance	Architecture comparison
η^2 **(UMAP)**	2D UMAP	ANOVA on projection	36.2%	5.6%	6.5×	Variance in visual space	Visualization only
Silhouette	2D UMAP	Cluster separation	0.095	0.005	20×	Within vs between cluster distance	Cluster quality
Bootstrap Variance	2D UMAP	Pairwise distances	103.7%	146.1%	1.41×‡	Within-group trajectory spread	Group cohesion

† **Important note:** Values shown are from post-scaffold analysis (Loops 1-7, n=20,000) demonstrating style dominance after reflection scaffolds are introduced. This targeted analysis isolates scaffold transformation effects (our primary finding: R4). A separate PERMANOVA on the full dataset (all loops, n=10,000) found style R^2=58.9% vs provider R^2=54.9%—this general analysis includes both Loop 0 (pre-scaffold) and Loops 1-7 (post-scaffold), showing overall geometry without isolating transformation. Both analyses confirm style dominance; we report post-scaffold values for mechanistic clarity.

‡ Inverted ratio (146.1/103.7) because lower variance indicates MORE cohesive groups. Style groups are more cohesive (103.7%) than provider groups (146.1%).

Key Insights from Reconciliation:

1. **Different metrics measure different properties**: R^2 and η^2 measure variance explained, silhouette measures cluster separation, bootstrap measures within-group spread

2. **All methods confirm style dominance** (except bootstrap, which measures cohesion inversely): Ratios range from 1.74× to 20×

3. **Cross-method consistency**: Both PERMANOVA (1.74×) and native ANOVA (3.63×) confirm style dominance through different mathematical approaches

4. **Bootstrap paradox resolved**: Provider shows higher bootstrap variance (146.1%) because provider groups are *less cohesive* (more spread out), which is consistent with weak provider clustering. *Higher variance = less cohesive, so style's lower value (103.7%) confirms its dominance.*

Usage Guidelines:

- **For primary claims about topology**: Use R^2 from PERMANOVA (rotation-invariant, native space)

- **For validation**: Cross-check with η^2 from native coordinate ANOVA (different method, confirms result)

- **For visualization**: Use η^2 from UMAP (matches what readers see in figures)

- **For cluster quality**: Use silhouette scores (intuitive separation metric)

- **For group cohesion**: Use bootstrap variance (lower = more cohesive)

All variance percentages in this paper reference this table unless otherwise specified.

Having established that style dominates manifold topology (R3), we now ask: *where does this style structure come from?* Does it emerge from the scaffolds themselves, or do scaffolds merely reorganize something already present? The following analysis reveals a surprising answer: scaffolds transform pre-existing category structure rather than creating new structure de novo.

3.4 R4: Category Structure Transformation (PERMANOVA Validated)

Key Finding: Scaffolds transform pre-existing category structure into style-driven organization, fundamentally reframing our understanding of prompt engineering.

Baseline Analysis (Loop 0, headless embeddings—seed prompts without model responses): Without reflection scaffolds, embeddings of seed prompts alone reveal strong inherent structure:

- **Category R^2 = 35.6%** (F=658.2, p<0.001, 999 permutations) - Massive category-based clustering

- **Style R^2 = 0.002%** (F=0.036, p=1.0) - Essentially zero style structure

- **Provider R^2 = 1.98%** (F=53.1, p<0.001) - Small but significant provider differences
- **N = 13,112 runs** (Loop 0 only, one embedding per run from seed prompts without model responses)

Post-scaffold Analysis (Loops 1-7): Introduction of reflection scaffolds dramatically redistributes variance:

- **Category R^2 = 9.3%** (F=189.7, p<0.001, 74% reduction from baseline)
- **Style R^2 = 17.2%** (F=421.3, p<0.001, +17.2 percentage points from near-zero baseline)
- **Provider R^2 = 9.9%** (F=242.8, p<0.001, 5× increase from baseline)
- **N = 91,784 total embeddings** (13,112 runs × 7 loops each with scaffolds; PERMANOVA uses stratified subsample n=20,000)

Statistical Validation:

- **Cross-method consistency**: Native coordinate ANOVA confirms transformation pattern
- **Robustness**: Results stable through 20+ bug fixes (v2→v3)
- **Independence**: Three separate PERMANOVAs, exploratory analysis (no Bonferroni correction)

Interpretation: The 74% reduction in category variance concurrent with style emergence (+17.2 percentage points from near-zero) demonstrates that scaffolds *redistribute* existing semantic organization rather than create structure de novo. Category clusters don't disappear—they transform into style-specific exploration patterns.

Note on ratio vs percentage point metrics: While style emergence from 0.002% to 17.2% represents an 8,600× ratio, we report absolute percentage point change (+17.2 pp) as the primary measure because the near-zero baseline (0.002%) makes ratio metrics sensitive to denominator precision. The absolute change reflects the substantive structural transformation.

Mechanistic Insight: To illustrate this pattern intuitively, scaffolds act like rearranging furniture in a room: the chairs, tables, and sofas already exist (semantic categories in embedding space), but their arrangement defines how the space functions. Scaffolds reposition what's already there, transforming topic-based organization into reflection-style-based clustering.

Causal caveat: While the correlation between scaffold introduction and variance redistribution is robust across validation methods (PERMANOVA, native ANOVA, bootstrap B=1000), establishing causality would require experimental manipulation (e.g., scaffold introduction at different loop points). We use "transform" as descriptive shorthand for this correlated variance shift. See Future Work for proposed ablations.

Validation: N=3 independent review (Claude Code, Codex gpt-5.1-codex-max, Gemini 2.5-pro) achieved unanimous agreement on the transformation hypothesis and statistical validity.

4 Discussion

Having established four empirical patterns—exploration-then-drift dynamics, coupled correlations, style dominance, and category transformation—we now examine their broader implications for understanding recursive reflection in language models.

4.1 Synthesis

Four findings converge on recursive reflection as *transformation* of existing structure:

1. **Dynamics (R1)**: Exploration-then-drift (not convergence) - variance grows 2.60× while shifts decline 71%
2. **Structure (R2)**: Coupled correlation blocks link local (shift↔effort) and global (variance↔distance) metrics
3. **Topology (R3)**: Style dominates provider in native 4096D space (1.74× ratio) and 2D projection (6.5×)
4. **Mechanism (R4)**: Scaffolds transform existing category structure (74% drop) into style structure (+17.2 pp emergence)[‡]

Scope: These characterize reflection dynamics (Loops 1-7). The initial response (Loop 0) reflects baseline behavior; subsequent reflection is shaped more by style than architecture.

4.2 Implications

For Reflection Research: Challenges assumption that self-observation leads to consensus. Reflection amplifies diversity while reducing volatility.

For Prompt Engineering: Style optimization appears to matter more than model selection (3.6× greater impact in native embedding space, 6.5× in 2D projection).

For Interpretability: Embedding trajectories provide measurable framework for characterizing introspection.

4.3 Limitations

1. **Anthropomorphism disclaimer**: 'Reflection' and 'introspection' are metaphors for recursive prompting, not claims about consciousness.
2. **UMAP dependency**: 6.5× ratio (η^2) measured in 2D UMAP (n_neighbors=15, min_dist=0.1). Different projections may yield different ratios. Native 4096D space shows 3.6× ratio, confirming robustness.
3. **E5-specific findings**: All primary results use E5-Mistral-7B embeddings. Multi-embedder sensitivity analysis (Appendix A) shows Style > Provider holds across MiniLM, MPNet, and E5-Large, confirming robustness within the E5/Sentence-Transformers family. Findings may not generalize to other embedding families (e.g., provider-native embeddings reverse the dominance pattern).

4. **Stabilization threshold**: 0.05 cosine shift is heuristic (chosen based on late-loop observed values).

5. **Incomplete dynamics**: Variance grows through Loop 7; may need 16-32 loops for asymptotic behavior.

6. **Provider ranking instability**: Absolute patterns robust; relative orderings are not (H3 failure).

7. **Geometric vs task performance**: We characterize embedding geometry, not downstream accuracy.

8. **Model heterogeneity**: Set includes reasoning models (10) and one non-reasoning model (gemini-2.5-flash-lite), which shows similar patterns.

9. **Prompt-text leakage (addressed through scaffold-free analysis)**: Initial concern that embeddings include scaffold text (reflection prompts), not just generated content, has been addressed. Scaffold-free analysis (Loop 0, "headless" embeddings—seed prompts without model responses, containing only category content) shows strong category structure (R^2 = 35.6%) with essentially zero style structure (R^2 = 0.002%), confirming scaffolds transform existing structure rather than introduce superficial prompt-driven separation.

10. **Temperature non-equivalence**: All models used T=0.7, but this temperature may not produce equivalent output distributions across different provider architectures. Normalizing stochasticity (e.g., calibrating temperatures to achieve similar entropy) or testing multiple temperatures would strengthen robustness claims.

11. **Category effects (analyzed)**: Category effects show striking transformation pattern - R^2 = 35.6% without scaffold (Loop 0) reduces to 9.3% with scaffold, while style R^2 increases from 0.002% to 17.2%. This suggests scaffolds redistribute semantic organization from topic-based to reflection-style-based clustering. Style×category interactions remain unexplored.

12. **PERMANOVA assumptions (acknowledged and validated)**: PERMANOVA's homogeneity of dispersion assumption is violated (PERMDISP $p<0.001$ for all factors), meaning R^2 values reflect both centroid location AND within-group spread. Rather than undermining findings, this dispersion heterogeneity *reinforces* the style dominance conclusion: style clusters are both more separated (higher R^2) AND more cohesive (tighter dispersion) than provider clusters. Additionally, potential pseudoreplication from within-run correlations was addressed via run-level PERMANOVA on Loop 7 only (N=13,112 independent observations), confirming style R^2=28.2% vs provider R^2=18.0% (ratio 1.57×, p=0.001). The consistent Style > Provider pattern across multiple validation methods demonstrates robustness to both assumption violations.

13. **Scaffold-token inclusion**: Loops 1-7 embeddings include scaffold instruction text alongside model responses. While this reflects real usage context (prompts and responses form a unified semantic context), style differences in Loops 1-7 may partly re-

flect scaffold vocabulary rather than purely model behavior. However, the Loop 0 baseline (no scaffold) establishes that category structure exists pre-scaffold (R^2=35.6%), and this structure transforms (74% reduction) with scaffold introduction—a pattern that holds regardless of scaffold-token contribution to Loops 1-7 clustering. A scaffold-stripped post-scaffold ablation would strengthen causal claims.

4.4 Methodology Transparency

Corrections made post-hoc:

- **Metric mismatch** (Obs 2): Curvature vs UMAP → Fixed, 100% validation
- **Aggregated testing** (Obs 5): Means vs per-provider → Fixed, 100% pass
- **NaN handling** (Obs 5 H3): Excluded vs failures → Fixed, honest 0% pass

4.5 Methodological Finding: Embedding Choice Determines Variance Attribution

Variance decomposition is embedding-dependent. Preliminary analysis with provider-native embeddings found architecture ∼90% vs style ∼10%. Universal E5 embeddings find the opposite: style $\eta^2 = 8.6\%$ vs provider $\eta^2 = 2.4\%$ in 384D PCA space (3.6× ratio, >99% of 4096D variance retained); style $\eta^2 = 36\%$ vs provider $\eta^2 = 6\%$ in 2D UMAP (6.5× ratio). Provider-native embeddings preserve architectural fingerprints; universal re-embedding reveals style as dominant organizer.

4.5.1 Validation in Native Embedding Space

Does Style > Provider hold in native E5-Mistral space (384D PCA of 4096D), or is it UMAP artifact? We computed η^2 across all 13,112 runs (104,896 total embeddings) with permutation testing (1000 iterations) and bootstrap CIs:

Table 4: Native Space Validation

Space	Style η^2	Provider η^2	Ratio	p-value
E5-Mistral (384D PCA)	**8.61%** [8.53, 8.70]	**2.37%** [2.35, 2.41]	**3.63**×	<0.001
UMAP 2D (projection)	36.2%	5.6%	6.5×	<0.001

Key findings:

1. **Style > Provider confirmed in 384D PCA space (>99% of 4096D variance)**: 3.63× ratio validates that dominance is a property of E5-Mistral embedding space, not UMAP artifact.
2. **UMAP amplifies non-uniformly**: Style amplification (4.2×: 8.61% → 36.2%) exceeds Provider amplification (2.4×: 2.37% → 5.6%). UMAP preserves local structure; style creates tight local clusters (more amplified).

3. **Statistical robustness**: Both effects highly significant with p = 0.001 across 1,000 permutations.

Note: These η^2 values (coordinates) are not directly comparable to other metrics; see Table 3 for complete reconciliation of all variance metrics.

E5-Specificity: These findings validate the Style > Provider pattern within E5 embedding space. The observation that provider-native embeddings reverse this pattern (§4.5) underscores that variance decomposition is measurement-frame dependent. We do not claim Style > Provider is a universal property of all embedding spaces—it is a robust property of E5 representations, validated across multiple E5 family members (see Appendix A).

4.6 Relation to Other Exploratory Findings

Provider Fingerprints vs Ranking Instability: Preliminary observations about provider "personalities" (e.g., Claude=exploratory, GPT=balanced) reflect central tendencies only; provider ranking order is not stable across reflection loops (ρ=-0.14, H3 failure), indicating no consistent "provider personality" in trajectory behavior. Fine-grained provider distinctions should not be interpreted as robust.

"No Decay" vs "Exploration-then-Drift": Turning-angle curvature (second derivative) is stable; first-derivative metrics (shift, effort) decline. Trajectory straightens locally while spreading globally.

4.7 Length Control Analysis

Response length varied 4.7× across providers (OpenAI 2,497 tokens/loop vs Anthropic 530). Analysis:

1. **Token-metric correlations are negative** (r ≈ -0.43): Longer responses show *less* drift.

2. **Token contribution is small**: +0.8-2.4% R^2 beyond style + provider.

3. **Patterns survive length control**: Style > provider persists in residuals.

Core findings are robust to verbosity.

4.8 Temporal Dynamics of Style Dominance

The 2.06× ratio masks temporal dynamics. Per-loop η^2 (UMAP coordinates):

Methodological note: The following per-loop η^2 values derive from 2D UMAP projections (exploratory analysis). While these temporal patterns are striking, formal per-loop PERMANOVA validation in native 4096D space would strengthen these claims—a promising direction for future work.

Temporal pattern: Style dominance appears immediate in UMAP projections (∼50-51% η^2 across all loops) and stable; provider signatures start near zero and grow to ∼9% by Loop 7.

Table 5: Temporal Dynamics

Loop	Style η^2	Provider η^2	Ratio
1	50.1%	0.6%	~80×
2	51.3%	5.0%	10×
7	51.8%	8.9%	6×

Remaining variance reflects interactions, content, and noise. Early chains (1-3 loops) are almost entirely style-driven in this projection; longer chains reveal accumulating provider fingerprints.

5 Conclusion

Through 13,112 independent runs and rigorous bootstrap validation (resampling at run-level to respect dependencies), we demonstrate that recursive reflection is fundamentally *transformation* of pre-existing semantic structure:

1. **Reflection is exploration-then-drift**: Global variance grows 2.60× while per-step shift declines by ~71%
2. **Trajectories organize into coupled correlation structure**: Two tight blocks (shift ↔ effort r=0.879, variance ↔ distance r=0.941) with strong cross-block coupling (variance ↔ shift r=0.683)
3. **Style dominates topology**: Rotation-invariant PERMANOVA shows style R^2 = 17.2% vs provider R^2 = 9.9% (ratio 1.74×, p=0.001)
4. **Scaffolds transform existing structure**: Category variance drops 74% (35.6% → 9.3%) while style emerges from near-zero to 17.2% (+17.2 percentage points), proving scaffolds redistribute rather than create structure de novo

These findings characterize recursive self-observation as *transformation* of inherent category organization into style-driven exploration patterns. Prompts don't create structure from nothing—they reorganize the semantic "furniture" already present in the embedding space. At the representational scale we probe, prompt engineering appears to be reorganization, not creation.

Future work: Extended loop sequences (16-32 loops), content category analysis, cross-model trajectory comparison, style transfer mid-sequence.

Scaffold-Free Ablation Analysis: A critical methodological question remains: our current analysis embeds full conversation context including scaffold instruction tokens. While Loop 0 (scaffold-free) provides a baseline showing zero style/provider structure, additional ablations are needed to fully disentangle scaffold text from model-generated content effects:

1. **Scaffold-free embeddings** (Loops 1-7): Re-embed with scaffold tokens masked to verify style dominance persists without prompt lexical cues.

2. **Output-only vs full-context comparison**: Compare PERMANOVA results between (a) full conversation embeddings and (b) response-only embeddings to quantify scaffold text contribution.

3. **Length-controlled analysis**: Normalize for context length growth (fixed-size sliding window or token budget) to rule out dilution effects in per-step shift decline patterns.

These ablations are planned for v2 and will strengthen the causal interpretation of the transformation-rather-than-creation finding.

Data availability: Full dataset and code available upon request.

6 Related Work

Our investigation of recursive self-reflection in language models sits at the intersection of four research areas: self-improvement methods, representation engineering, dimensionality reduction for LLM analysis, and prompt engineering.

6.1 Self-Reflection and Self-Improvement

Self-Refine [7], **Reflexion** [10], **ReAct** [15], and **Tree of Thoughts** [14] assume reflection leads to *convergence*. We observe geometric *divergence*, suggesting reflection is exploration at the embedding level.

Recent surveys of multi-step reasoning [9, 6] document that multi-turn processes often explore rather than strictly converge. Our geometric divergence finding (variance grows 2.60× while shifts decline 71%) provides quantitative evidence for what these surveys observe behaviorally: reflection is exploration, not refinement.

6.2 Representation Engineering

Representation Engineering [16] and work on world models [4] probe static representations. We extend to *dynamic* analysis: tracking representations across recursive iterations.

6.3 Dimensionality Reduction

UMAP [8] and t-SNE visualize embeddings. We apply UMAP to *multi-loop trajectories*, revealing exploration-then-drift and style-dependent signatures invisible in single-point clustering.

Recent work on UMAP force shapes and initialization sensitivity [5] shows that UMAP can amplify or suppress structure depending on hyperparameters and local density. Our finding that UMAP amplifies style structure more than provider structure (4.2× vs 2.4× amplification from 4096D to 2D) aligns with UMAP's sensitivity to tight local clusters, which style creates more strongly than provider in E5 space.

6.4 Prompt Engineering

Chain-of-Thought [13] and Self-Consistency [12] showed prompting affects capabilities. Our finding that style explains 3.6× more variance than architecture (native 4096D space) provides geometric evidence for what prompting literature demonstrates behaviorally.

Recent work on delimiter-format brittleness [11] shows that minor prompt formatting changes can flip model rankings and behaviors. Our finding that style dominates geometry in E5 space (6.5× style/provider ratio in 2D UMAP, 3.6× in native 4096D space) suggests that prompt scaffolding (style) is as consequential as model identity (provider) for shaping iterative reflection geometry, with immediate implications for evaluation design and deployment.

6.5 Positioning

Novel contributions:

1. **Trajectory analysis**: Treating reflection as paths reveals dynamics invisible to single-point analysis

2. **Quantified style dominance**: First geometric measurement (style $\eta^2 = 8.6\%$ vs provider $\eta^2 = 2.4\%$ in native 4096D E5-Mistral space; 36% vs 6% in 2D UMAP projection)

3. **Bootstrap-validated topology**: Rigorous per-provider validation with transparent failure reporting

Acknowledgments

Research conducted independently by twin brothers in Alaska, without institutional support. Methodology corrections applied transparently.

Principles: This work embodies four transparency principles:

1. **Honest failure reporting**: H3 rejection (provider ranking instability) reported prominently, not buried

2. **Correctness over convenience**: Fixed methodology issues with full documentation (§2.5)

3. **Evidence through bootstrap**: All retained primary findings pass B=1000 validation at p<0.001 (H3 rejected)

4. **Accountability in corrections**: N=3 review trail documented, commit history preserved

AI Assistance: Reviewed with multiple AI systems for technical accuracy. Final responsibility rests with human authors.

References

[1] M. J. Anderson. A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46, 2001. doi: 10.1111/j.1442-9993.2001.01070.pp.x.

[2] M. J. Anderson. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, 62(1):245–253, 2006. doi: 10.1111/j.1541-0420.2005.00440.x.

[3] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition, 1988.

[4] W. Gurnee and M. Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[5] M. T. Islam and P. Fleischer. The shape of attraction in umap: Exploring the embedding forces in dimensionality reduction. *arXiv preprint arXiv:2503.09101*, 2025.

[6] Y. Li et al. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*, 2025.

[7] A. Madaan et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[8] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[9] A. Plaat, W. Kosters, and M. Preuss. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

[10] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[11] J. Su, T. Zhang, K. Ullrich, L. Bottou, and M. Ibrahim. A single character can make or break your llm evals. *arXiv preprint arXiv:2510.05152*, 2025.

[12] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

[13] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

[14] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[15] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

[16] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Multi-Embedder Validation

A.1 Shift/Variance Dynamics (R1)

Table 6: Multi-Embedder Dynamics

Embedder	Dim	Shift Decline	Variance Growth
MiniLM	384	40%	3.6×
MPNet	768	49%	3.9×
E5-Large	1024	52%	2.60×
E5-Mistral[†]	4096	**71%**	**2.60**×

[†] Primary embedder. All show shift decline by 40-71% and variance growth of 2.60-3.9×, confirming R1 dynamics are embedder-agnostic.

A.2 Style/Provider η^2 Topology (R3)

Table 7: Multi-Embedder Topology

Embedder	Dimensions	Style η^2	Provider η^2	Ratio	Convergence
MiniLM-L6-v2	384	28.18%	7.60%	**3.71**×	✓10.1%
MPNet-base-v2	768	22.89%	9.34%	**2.45**×	✓6.4%
E5-Large-v2	1024	24.69%	14.29%	**1.73**×	✓1.7%
E5-Mistral-7B	4096	30.42%	15.99%	**1.90**×	✓0.1%

Inverse Dimensionality-Style Relationship: Lower-dimensional embeddings show stronger style dominance (384D: 3.71×) compared to higher-dimensional embeddings (1024D: 1.73×), with E5-Mistral (4096D: 1.90×) showing deviation from perfect monotonicity.

Independent Computational Validation: To validate statistical interpretations, we employed three independent AI systems (Claude Code, Codex, Gemini) as computational reviewers—a form of automated methodological review. This approach tests whether conclusions (e.g., style dominance pattern) are derivable from statistical results alone, without coordination between reviewers. The inverse dimensionality-style correlation achieved **Statistical Consensus (3/3)** on pattern existence but **Interpretative Divergence (0/3**

agreement) on underlying mechanisms. This divergence is methodologically informative: the pattern's existence is statistically robust, but mechanistic interpretation requires domain expertise beyond statistical analysis.

High-Priority Concerns: Both Codex and Gemini raised critical issues that must be addressed before this finding can be considered fully validated:

1. **PCA Projection Bias**: Using a fixed K=50 for PCA may bias comparisons across embeddings of different dimensions.

2. **Model Architecture Confound**: The analysis does not disentangle the effects of dimensionality from the models' underlying architecture and training objectives.

Convergence: All four embedders passed ANOVA convergence validation (<20% difference between ANOVA on coordinates vs distances), confirming the style > provider pattern is not a methodological artifact.

B Response Length by Provider

Table 8: Response Length Statistics

Provider	Outputs	Mean Tokens	Max	>8,192
Anthropic	28,656	530	15,893	8
OpenAI	19,072	2,497	16,160	314
Google	19,088	1,558	16,457	21
xAI	19,008	1,456	14,234	18
DeepSeek	9,536	541	3,874	0
Moonshot	9,536	1,872	12,456	6
Total	**104,896**	**1,419**	-	**367**

Despite 4.7× variation, patterns are robust to length control (§4.7).

C Authoritative Findings Reference (Summary)

Full document: *`docs/findings/authoritative-findings-v2.md`*

C.1 Dataset Summary

C.2 Core Results Mapping

C.3 Validation Completeness

D Methodological Rigor & Transparency (Summary)

Full document: *`docs/addendum-methodology.md`*

Table 9: Dataset Summary

Metric	Value	Source
Experiments	13,112	`full-scale-2025-11-20-v2` batch
Embeddings	104,896	Complete trajectories (13,112 × 8 loops)
Providers	6	Anthropic, Google, OpenAI, DeepSeek, Moonshot, xAI
Models	11	See authoritative findings for breakdown
Styles	8	7 scaffolds + 1 baseline (none)
Categories	12	Various semantic domains
Completeness	100%	All experiments have complete 8-loop trajectories

Table 10: Core Results

ID	Finding Title	Key Metric	Validation
R1	Exploration-Then-Drift Dynamics	71% shift decline, 2.6× variance growth	Bootstrap B=1000, p<0.001
R2	Two-Dimensional Correlation Structure	r=0.88 (Shift↔Effort), r=0.94 (Var↔Dist)	Pearson correlation
R3	Style-Dominant Manifold Topology	R^2 17.2% vs 9.9% (1.74×), η^2 8.61% vs 2.37% (3.63×)	PERMANOVA, native ANOVA
R4	Category Structure Transformation	Category 35.6%→9.3% (F=189.7), Style 0.002%→17.2% (F=421.3)	PERMANOVA, 999 permutations

Table 11: Validation Status

Type	Status	Details
Data Completeness	✓	13,112/13,112 (100%)
N=3 Verification	✓	Codex + Gemini + Claude unanimous
Bootstrap (B=1000)	✓	All findings validated
PERMANOVA	✓	Rotation-invariant validation
Run-Level PERMANOVA	✓	Loop 7 only, N=13,112 independent: Style R^2=28.2% vs Provider R^2=18.0%
PERMDISP	✓	Dispersion homogeneity test (Anderson 2006): F=487.75 (style), 104.92 (provider), 42.02 (category), all p<0.001
Scaffold-Free	✓	Loop 0: Category η^2=35.6%, Style η^2=0.002% (pre-scaffold)

D.1 Scientific Approach

Experimental Design:

- **Configuration-as-code paradigm** ensures reproducibility (`cmd/experiment/batches.go` as immutable source)
- **Progressive validation:** Pilots (N=660) → Full-scale (N=13,112)
- **Factorial design:** 6 providers × 8 styles × 12 categories with controls
- **Deterministic execution:** Unique IDs from parameter hash

Statistical Validation:

- Bootstrap validation (B=1000, $p<0.001$) rather than parametric assumptions
- Random seeds controlled for reproducibility
- Confidence intervals transparently reported

Confound Handling:

- ✓Multi-embedder validation (4 models: 384D-4096D)
- ✓Native space analysis (pre-UMAP validation)
- ✓Per-provider testing (patterns in ≥5/6 providers)
- ✓Scaffold prompt leakage addressed (Loop 0 baseline confirms category structure preexists at 35.6%)
- △ Single embedding family limitation (E5 only)
- △ Short trajectories (8 loops maximum)

D.2 Implementation Quality

Strengths:

- Modular Go pipeline (`experiment`, `validate`, `visualize`)
- Idempotent operations for safe retries
- Type-safe configuration with registries
- Rich JSONL provenance logging

Areas for Improvement:

- Command-layer testing coverage (2/29 files)
- Manual environment loading requirement
- Memory requirements (96GB for E5-Mistral-7B)

D.3 Documentation & Transparency

- 80+ implementation plans in `docs/plans/`
- 113+ review files with N=3 validation
- All corrections documented with commit trail
- Bootstrap/validation artifacts preserved

Assessment Method: N=3 Independent Review (Gemini, Codex, Claude) with unanimous agreement on core methodology validity.