SPECTRAL INVARIANTS OF REASONING TRAJECTORIES: A KOOPMAN-DMD DYNAMICAL VIEW ON LLM INTERNAL LOGIC GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models produce multi-step natural-language reasoning chains, yet little is known about whether distinct logical inference types induce separable intrinsic geometric signatures within latent hidden-state dynamics. Moving beyond static-layer probing, we treat cumulative reasoning hidden-state sequences as discrete phase-space trajectories and adopt Koopman operator theory together with Exact Dynamic Mode Decomposition (DMD) to extract complex-valued spectral fingerprints of internal reasoning dynamics. A major practical obstacle is that DMD eigenvalue sets from different reasoning traces vary in cardinality; naive zero-padding introduces artificial mass at the origin and corrupts spectral comparisons. To address this, we develop a variable-rank W_2 optimal-transport metric for unaligned complex eigenvalue clouds, avoiding zero-padding distortions. We state and empirically investigate the *Spectral Invariant Hypothesis* (SIH): spectral fingerprints from identical logic types exhibit smaller intra-type W_2 distances than inter-type distances. Evaluated on controlled logical reasoning benchmarks across multiple open-weight models, our results confirm the hypothesis and reveal non-monotonic scaling of dynamical manifold structure with model size. Our dynamical-systems perspective provides a probe-free diagnostic paradigm for analyzing LLM reasoning mechanisms.

1 INTRODUCTION

Large language models (LLMs) solve complex logical deduction tasks by generating multi-step derivation traces in natural language (Wei et al., 2022; Kojima et al., 2022; Chen et al., 2026; Nguyen et al., 2025; Lucas et al., 2026). While these verbal chains of thought make intermediate reasoning steps explicit at the token level, a fundamental question remains unaddressed: does the generation of distinct logical derivation classes induce geometrically distinguishable flow signatures within the model's internal latent space?

Answering this question carries significant practical and theoretical stakes. If internal reasoning trajectories follow discoverable geometric manifolds, those structures can serve as probe-free diagnostics for mechanistic interpretability—allowing us to monitor, audit, and analyze the internal computation of models without task-specific tuning heads or external supervision. Standard probing paradigms (Hewitt & Manning, 2019; Tenney et al., 2019; Belinkov, 2022; Kissane et al., 2025; Servedio et al., 2025; Zheng et al., 2026) predominantly operate on static representations sampled at isolated individual layers. However, static snapshots inherently fail to capture the temporal continuity and state evolution that govern sequential inference. Recent investigations into transformer representation dynamics indicate that internal residual streams exhibit quasi-continuous flow behavior, displaying high adjacent-layer similarity and low-rank transition maps across intermediate depth and temporal rollouts (Ma et al., 2026; Lawson et al., 2025; Jiang et al., 2025; Fernando & Guitchounts, 2025).

A dynamical systems perspective. Rather than treating multi-step reasoning as a concatenation of isolated token representations, we view the forward rollout of hidden states as a discrete trajectory traversing an underlying phase space. Formally, let the contextual hidden states evolving across the inference sequence be viewed as state snapshots of a nonlinear dynamical system governed by an

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

unknown continuous or discrete evolution rule. Directly analyzing such nonlinear flows is analytically challenging; however, Koopman operator theory (Koopman, 1931; Mezić, 2005) provides an exact linear, infinite-dimensional representation of nonlinear autonomous dynamics through the lens of measurement functions (observables). To render this framework computationally tractable over high-dimensional neural states, we employ Exact Dynamic Mode Decomposition (DMD) (Schmid, 2010; Tu, 2013) as a data-driven finite-rank approximation of the underlying Koopman operator. Through this lens, the latent reasoning path is decomposed into characteristic spectral modes and complex eigenvalues, which provide a compact spectral characterization of temporal patterns unfolding across the reasoning process.

Comparing these extracted spectral signatures across disparate reasoning traces, however, poses a fundamental geometric challenge. Because each instance generates a variable-length chain of thought, the resulting eigenvalue point clouds lack a canonical ordering and vary in cardinality across sample trajectories. Standard vector-space distances or naive element-wise alignments fail in this setting, as they either impose arbitrary pairings or introduce artificial coordinate dependencies. To overcome this, we turn to the optimal transport framework (Villani et al., 2009; Gabriel & Marco, 2019) and model each eigenvalue collection as a discrete probability measure in the complex plane. Specifically, the 2-Wasserstein metric (W_2) measures the minimal transport cost required to morph one spectral point cloud into another under Euclidean ground distance (Kantorovich, 1960; Gabriel & Marco, 2019). In our setting, reasoning traces vary in length and dynamic complexity, yielding Exact DMD point clouds with differing truncation ranks r. Standard Wasserstein distances require arbitrary zero-padding to handle varying cardinalities, which artificially pools mass at the origin and distorts spectral drift. Our variable-rank W_2 formulation resolves this cardinality mismatch directly, aligning spectral footprints in $\mathbb{C}$ while preserving active mode densities to isolate true geometric divergence across distinct logic types. We formalize this inquiry through the *Spectral Invariant Hypothesis* (SIH):

Hypothesis 1 (SIH)**.** *For each carrier model, the spectral fingerprint $\mathcal{F}(\alpha) = \{\lambda_i\}_i \in \mathbb{C}^r$ obtained by Exact Dynamic Mode Decomposition (*DMD*) of the context-cumulative reasoning trajectory exhibits W_2 separation by logic type: intra-logic W_2 distances are strictly smaller than inter-logic W_2 distances, i.e., $W_2^{\text{intra}} \ll W_2^{\text{inter}}$.*

The SIH hypothesis fundamentally makes a claim about the intrinsic geometry of hidden-state reasoning trajectories inside large-language models. Instead of evaluating model outputs, classification decision boundaries, or robustness to input perturbations, it describes how dynamical patterns from reasoning processes cluster or spread out within the metric space formed by DMD spectral fingerprints. Measuring structural differences via optimal-transport distance offers an intuitive geometric way to formalize this idea. The W_2 separation quantifies how dissimilar these spectral manifolds are, and it works without relying on arbitrary coordinate choices, input-output dimension scaling, or unwanted metric distortions. Our contributions are as follows:

- We recast LLM multi-step reasoning traces as phase-space trajectories and introduce variable-rank Exact DMD on context-cumulative state matrices to extract coordinate-free spectral fingerprints.
- We formulate a variable-rank W_2 metric over finite eigenvalue point clouds in $\mathbb{C}$ for comparing dominant dynamical modes across heterogeneous-rank trajectories, alongside theoretical analysis of its topological behaviour and error bounds versus zero-padded baselines.
- We empirically validate the SIH across multiple contemporary LLM families, demonstrating clear geometric separation, and distinctive patterns of cross-model representation alignment across open-weight architectures.

2 Related Work

Recent research has investigated the intrinsic geometric properties of large-language-model (LLM) internal representations, demonstrating that token embeddings reside on low-dimensional manifolds instead of occupying the full high-dimensional ambient space (Modell et al., 2025; Kataiwa et al., 2025; Viswanathan et al., 2025). Across model layers, the intrinsic dimensionality of these representations dynamically expands and contracts throughout sequential decision-making processes (Joshi et al., 2026), and such underlying information-geometric structure can be exploited for

minimal-perturbation manifold steering interventions (Picozzi, 2026). These manifold-constrained representation spaces support semantic alignment: multilingual models shift non-dominant language reasoning toward a dominant-language geometric subspace (Zhang et al., 2025). Beyond Euclidean geometry, hyperbolic manifolds better model hierarchical syntax and semantic entailment in LLM activations (Yang et al., 2026).

Multi-step logical deduction has recently been framed as a continuous geometric flow evolving through representation space, guided by underlying semantic constraints and logical invariants (Zhou et al., 2026). Examining the topology and kinematics of these latent trajectories offers powerful diagnostic capabilities: reasoning failures and hallucinations may be characterized via geometric metrics including manifold deviation, structural curvature, and arrested displacement along inference manifolds (Li et al., 2026; Jiang et al., 2026b). These spatial properties also shed light on the internal mechanisms behind model self-correction, supporting targeted rollback interventions that dynamically detect high-uncertainty units via real-time perplexity during generation and regenerate them before errors propagate (Gao et al., 2026), and offer a foundation for probing emergent deceptive model behaviors, revealing that leading LLMs exhibit spontaneous deception tendencies even under entirely benign prompts (Wu et al., 2026).

Finally, Chain-of-Thought (CoT) reasoning has also been investigated from structural and geometric perspectives. Explicit sequential CoT vs. implicit latent thought reveals depth-recurrent benefits: looped Transformers enable parallel latent computation, more efficient than sequential CoT (Saunshi et al., 2025); test-time block iteration yields large reasoning gains without specialized CoT data (Geiping et al., 2026). Beyond theoretical analysis, structural-algebraic abstractions have been applied to refine CoT reasoning paths (Ranaldi et al., 2025). Such paradigms are further extending toward multimodal settings, enabling multimodal chain-of-thought reasoning (Jiang et al., 2026a) and 3D-informed compound design for large multimodal models (Ma et al., 2025).

3 METHOD

3.1 CONTEXT-CUMULATIVE TRAJECTORY CONSTRUCTION

Let $\tau = (s_1, s_2, \ldots, s_T)$ be a reasoning trace consisting of T step-level segments for solving a logic problem. At each step t, we form the *context-cumulative prefix* $(s_1, \ldots, s_t)$ and run a full forward pass through the LLM. We extract the hidden states of the newly appended tokens (i.e., those belonging to s_t) and mean-pool them to obtain a single state vector:

$$\mathbf{y}_t = \mathrm{MeanPool}\big(\mathrm{HiddenState}(s_1, \ldots, s_t)[s_t]\big) \in \mathbb{R}^d, \tag{1}$$

where d is the model's hidden dimension. Stacking these vectors column-wise yields the *trajectory matrix*

$$Y = [\mathbf{y}_1,\ \mathbf{y}_2,\ \ldots,\ \mathbf{y}_T] \in \mathbb{R}^{d \times T}. \tag{2}$$

This context-cumulative construction ensures that $\mathbf{y}_t$ encodes the model's representation of step s_t conditioned on the full prefix, not just the local token representation in isolation.

3.2 EXACT DYNAMIC MODE DECOMPOSITION

We model the evolution of hidden-state trajectories as a linear dynamical system in a low-dimensional subspace. Dynamic Mode Decomposition (DMD) (Schmid, 2010; Tu, 2013) is a data-driven method that extracts the dominant spatiotemporal modes from sequential data by approximating the best-fit linear operator advancing the system one step forward. This yields a compact spectral representation of the underlying dynamics and is therefore well-suited to capturing the intrinsic low-dimensional structure of LLM reasoning trajectories.

Given Y, we define the snapshot pairs

$$Y_1 = [\mathbf{y}_1, \ldots, \mathbf{y}_{T-1}], \quad Y_2 = [\mathbf{y}_2, \ldots, \mathbf{y}_T]. \tag{3}$$

We seek a linear operator A that advances the hidden-state trajectory one step forward in time, i.e., $\mathbf{y}_{t+1} \approx A\mathbf{y}_t$, so that the snapshot pairs approximately satisfy $Y_2 \approx AY_1$. The motivation for this linear ansatz comes from Koopman operator theory (Koopman, 1931; Mezić, 2005), which states that any nonlinear dynamical system can be lifted to an infinite-dimensional linear operator acting on

observables; DMD provides a finite-dimensional data-driven approximation of this operator restricted to the subspace spanned by the observed snapshots.

Because A may be high-dimensional and the trajectories exhibit strong intrinsic low-rank structure, we follow the Exact DMD formulation of Schmid (2010); Tu (2013) and approximate the operator on the dominant left-singular subspace of Y_1. Computing the SVD $Y_1 = U\Sigma V^\top$ and projecting onto the leading r modes, we obtain the *companion matrix*

$$\boxed{\tilde{A} = U^\top Y_2 V \Sigma^{-1}} \in \mathbb{R}^{r\times r}, \tag{4}$$

where r is the effective rank determined per trajectory. The spectrum of $\tilde{A}$ inherits the dominant spectral properties of the underlying infinite-dimensional Koopman operator while remaining computationally tractable. The eigenvalues $\{\lambda_i\}_{i=1}^r \subset \mathbb{C}$ of $\tilde{A}$ are the *dynamic modes*; they constitute the spectral fingerprint

$$\mathcal{F}(\alpha) = \{\lambda_i\}_{i=1}^r \subset \mathbb{C}^r. \tag{5}$$

Each eigenvalue λ_i encodes the temporal frequency (its argument) and growth or decay rate (its modulus) of a corresponding coherent mode in the reasoning trajectory, so that $\mathcal{F}(\alpha)$ provides a compact, interpretable summary of the underlying dynamics.

Variable effective rank. Fixed-rank truncation enforces identical dimensionality across all trajectories, conflating reasoning sequences with distinct intrinsic dynamical dimensionality. We instead compute a trajectory-wise effective rank from singular values of Y_1:

$$r = \text{eff_rank}(Y_1;\, \varepsilon) = \max\{k : \sigma_k \geq \varepsilon \cdot \sigma_1\}, \tag{6}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots$ are the singular values of Y_1 sorted in descending order, and ε denotes a relative threshold. This adaptive selection retains only dynamically significant modes above the threshold, assigning a tailored rank to each trajectory that reflects its intrinsic dynamical complexity.

3.3 Variable-Rank Wasserstein-2 Distance

To compare spectral fingerprints, we treat each set of eigenvalues as a discrete distribution in the complex plane and measure their similarity using the Wasserstein-2 distance. Unlike Euclidean distances between summary statistics, W_2 respects the geometry of the complex plane and provides a natural optimal-transport-based matching between modes, making it robust to permutations and variations in the number of eigenvalues.

Given two spectral fingerprints $\mathcal{F}(\alpha_a)$ and $\mathcal{F}(\alpha_b)$ with ranks r_a and r_b respectively, we define the **variable-rank** W_2 distance as follows. Let $k = \min(r_a, r_b)$. Sort both sets of eigenvalues by magnitude in descending order and retain the top-k elements from each. Apply the Hungarian assignment algorithm (Kuhn, 1955) to find the optimal bijection π^* minimizing total squared distance in $\mathbb{C}$:

$$W_2\big(\mathcal{F}(\alpha_a),\, \mathcal{F}(\alpha_b)\big)^2 = \frac{1}{k} \min_{\pi} \sum_{i=1}^{k} \big|\lambda_i^{(a)} - \lambda_{\pi(i)}^{(b)}\big|^2, \tag{7}$$

where the $\min$ is over all bijections $\pi : [k] \to [k]$. This variable-rank formulation compares only the dominant modes and eliminates any influence from zero-padding, which can artificially inflate similarity. We then compute intra-class distances (between fingerprints of the same logic type) and inter-class distances (between different logic types), and report their ratio.

4 Theoretical Analysis

As described in Section 3.3, we compare two strategies for handling variable-rank spectral fingerprints:

- **Variable-rank (ours):** truncate both spectra to $k = \min(r_a, r_b)$ eigenvalues sorted by $|\lambda|$ descending, then apply Hungarian matching.
- **Zero-padded:** extend the shorter spectrum to the longer rank $k = \max(r_a, r_b)$ by appending zeros, then apply Hungarian matching over all k elements.

We formalize the relationship between the distances produced by these two methods.

Proposition 1 (Upper Bound on Zero-Padded Penalty)**.** *Let $\mathcal{F}_a$ and $\mathcal{F}_b$ be two spectral fingerprints with effective ranks r_a and r_b. Assume without loss of generality that $r_a \geq r_b$, so that $K = \max(r_a, r_b) = r_a$ and $k = \min(r_a, r_b) = r_b$. Let W_2^{var} denote the variable-rank distance and W_2^{pad} denote the zero-padded distance. The total unnormalized transport cost of the zero-padded method is upper-bounded by the unnormalized variable-rank cost plus the squared norm of the truncated tail of the higher-rank trajectory:*

$$K\big(W_2^{\mathrm{pad}}\big)^2 \leq k\big(W_2^{\mathrm{var}}\big)^2 + \sum_{i=k+1}^{K} \big|\lambda_i^{(a)}\big|^2. \tag{8}$$

This theoretical result further justifies our choice of the variable-rank formulation, which avoids injecting artificial zero entries and their associated distortion entirely. Proofs and additional theoretical propositions are provided in Appendix D.

5 Experiments

5.1 Dataset and Setup

Dataset. We use Reasoning-Flow (Zhou et al., 2026), a publicly released benchmark consisting of 30 distinct abstract logical structures, where each structure corresponds to reasoning traces of 8–16 steps. For each logical structure, the benchmark preserves identical logical scaffolding while systematically varying superficial characteristics: every structure is instantiated across 20 topical domains and four languages (English, Chinese, German, Japanese). This controlled design yields a total of 2,430 reasoning sequences, which separate underlying logical form from surface-level semantics by holding the deductive template fixed while changing topical domain and linguistic realization.

Trajectory extraction. All five models are operated in pure inference mode, without fine-tuning or custom prompt engineering beyond standard instruction formatting. For each reasoning problem, we generate step-wise reasoning traces and extract context-cumulative trajectory matrices following Section 3.1. We apply exact DMD to each trajectory, retaining modes via a singular-value magnitude threshold $\varepsilon = 10^{-10}$ instead of fixed-rank truncation or zero-padding. This yields a variable-length spectral fingerprint per trace, where the effective rank is determined adaptively from the snapshot spectrum. No gradients are computed during trajectory extraction; cached fingerprints are reused across all downstream ablation and statistical analyses.

5.2 Main Result: W_2 Separation

Table 1: **Spectral fingerprint W_2 separation.** Intra-logic W_2: mean pairwise distance between fingerprints of the *same* logic type. Inter-logic W_2: mean pairwise distance between fingerprints of *different* logic types. Ratio: inter/intra. 95% CI: bootstrap confidence interval on the ratio (1000 iterations).

Model	W_2^{intra}	W_2^{inter}	**Ratio**	**95% CI**
Qwen3-0.6B	0.1515	0.3776	2.49	$[2.03, 2.58]$
Qwen3-1.7B	0.1465	0.3699	2.53	$[2.08, 2.58]$
Qwen3-4B	0.1442	0.3685	2.56	$[2.07, 2.64]$
Qwen3-8B	0.1581	0.3754	2.37	$[1.90, 2.53]$
Meta-Llama-3.1-8B	0.1409	0.3759	2.67	$[2.18, 2.74]$

Table 1 summarizes intra-logic and inter-logic W_2 distances, alongside their inter/intra ratio, across all five evaluated models. For every model, the mean inter-logic W_2 distance is substantially larger than the intra-logic counterpart, yielding ratios ranging from $2.37\times$ to $2.67\times$. This separation effect is robust and consistent across model sizes and architectures. The smallest ratio (2.37, Qwen3-8B) remains markedly greater than 1, while the largest value (2.67, Meta-Llama-3.1-8B) indicates

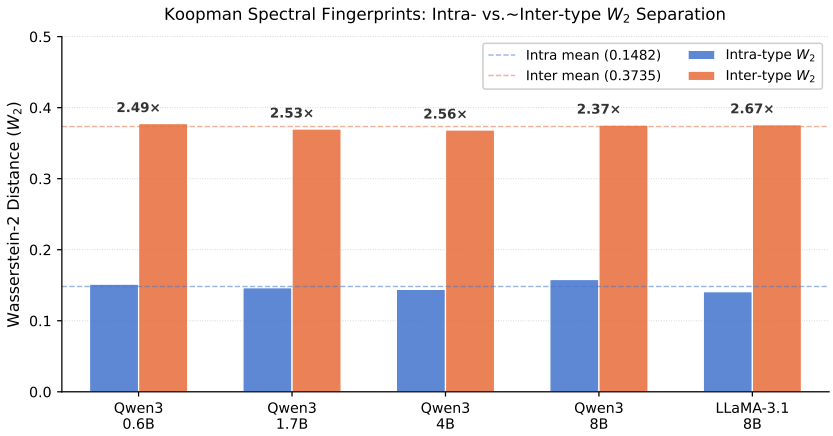


Figure 1: W_2 **separation across five LLMs.** Grouped bars show intra-logic (blue) and inter-logic (orange) W_2 distances per model. Numbers above the orange bars are inter/intra ratios.

Table 2: **Permutation test results.** The observed ratio far exceeds the null distribution for every model. $p = 0$ means zero of 500 permutations achieved the observed ratio.

Model	$\mathcal{T}_{\text{obs}}$	Null mean	Null pct$_{99}$	p
Qwen3-0.6B	2.49	1.000	1.011	$0/500$
Qwen3-1.7B	2.53	1.000	1.010	$0/500$
Qwen3-4B	2.56	1.000	1.011	$0/500$
Qwen3-8B	2.37	1.000	1.011	$0/500$
Meta-Llama-3.1-8B	2.67	1.000	1.010	$0/500$

stronger spectral separation in this larger instruction-tuned model. Figure 1 visually reinforces this pattern: the orange bars for inter-logic distances consistently sit far above the blue intra-logic bars for all five LLMs.

We validate statistical reliability via logic-type-level bootstrap resampling (1000 iterations, resampling the 30 logic types with replacement). The resulting 95% bootstrap confidence intervals all have lower bounds above 1.90, meaning the separation signal is unlikely to arise from sampling noise. Notably, this pattern holds universally across our dataset: for all 30 logical structures, the average inter-logic W_2 distance exceeds the average intra-logic distance ($30/30$, fraction $= 1.00$). This uniform trend suggests that the spectral fingerprints derived from DMD capture intrinsic differences between reasoning dynamics governed by distinct logical forms, rather than being driven by a small subset of outlier logic types.

5.3 Permutation Test

To assess statistical significance, we run a permutation test against the null hypothesis $\mathcal{H}_0$ that logic-type labels carry no information about spectral fingerprint geometry. Under $\mathcal{H}_0$, we randomly permute logic-type assignments across all traces and recompute the inter/intra W_2 ratio for each of $n_{\text{perm}} = 500$ permutations. Each permutation trial uses all n_{sample} traces per logic class, precomputing pairwise distances once and re-aggregating under permuted labels. We report the observed test statistic $\mathcal{T}_{\text{obs}}$ (the true ratio), the null mean, the null 99th percentile, and the empirical p-value $p = |\{T_{\text{perm}} \geq \mathcal{T}_{\text{obs}}\}|/n_{\text{perm}}$.

Table 2 shows that the observed ratio exceeds the 99th percentile of the null distribution by a factor of more than $2\times$ for every model (null pct$_{99} \approx 1.011$; observed 2.37–2.67$\times$). All five p-values are $0/500$: no permutation in 500 trials achieves the observed ratio, establishing $p < 0.002$ per model, strongly rejecting the null hypothesis and confirming that the spectral separation signal cannot be explained by random label assignment.

Table 3: **Spectral clustering of fingerprint manifold.** Silhouette score at the best k. Qwen3-0.6B, 1.7B, and Meta-Llama-3.1-8B prefer $k = 2$; Qwen3-4B and 8B prefer $k = 3$.

Model	Best k	sil
Qwen3-0.6B	2	0.427
Qwen3-1.7B	2	0.457
Qwen3-4B	3	0.355
Qwen3-8B	3	0.333
Meta-Llama-3.1-8B	2	0.405

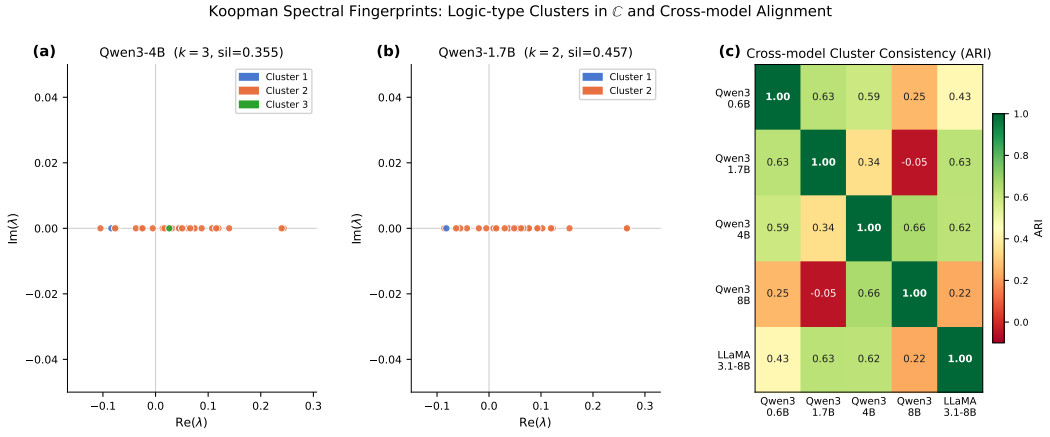
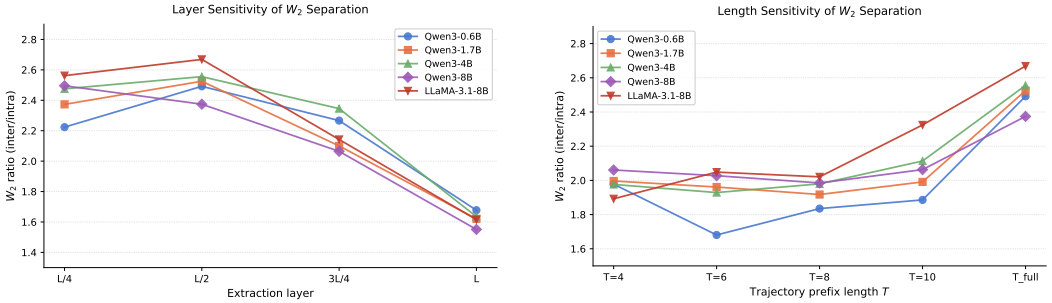


Figure 2: Koopman spectral clustering results. (a,b): Each dot denotes one logic-type centroid in the complex plane, colored by k-means cluster assignment. Visualization centroids are computed from *all* DMD eigenvalues across trajectories of a given logic type, by taking the median of real and imaginary components. Panel titles report the selected number of clusters k and silhouette score. Panel (a) shows Qwen3-4B and (b) shows Qwen3-1.7B. (c) Adjusted Rand-index (ARI) heatmap quantifies pairwise agreement of cluster partitions across models; high ARI indicates similar grouping of logic types in spectral space.

5.4 Spectral Clustering

We perform k-means clustering over per-logic spectral-fingerprint centroids, where each centroid corresponds to the median-matched eigenvalue cloud for one logic type in $\mathbb{C}$. For each model, we select the number of clusters k by maximizing the silhouette score $\mathrm{sil}(k)$ across the search range $k \in \{2, \ldots, 10\}$. Silhouette scores range from 0.33 to 0.46, providing evidence for moderately well-separated, geometrically meaningful clusters. Qwen3-0.6B and Qwen3-1.7B select $k = 2$, pointing to a coarse binary partition over reasoning dynamics. By contrast, Qwen3-4B and Qwen3-8B favor $k = 3$, indicating that larger-capacity Qwen3 models resolve finer-grained dynamical structure among logic types. Meta-Llama-3.1-8B also picks $k = 2$, yet its cluster assignments differ markedly from small Qwen3 models, which we attribute to its divergent training setup and model lineage. Figures 2 (a) and (b) show the results for Qwen3-4B and Qwen3-1.7B.

Cross-model cluster alignment. To measure whether different models group the 30 logic types in analogous ways, we compute the adjusted Rand index (ARI) between pairwise model cluster assignments in Figure 2 (c). Higher ARI values indicate greater agreement on how logic types are partitioned in spectral space. Within the Qwen3 family, the 4B/8B pair achieves the highest alignment (ARI $= 0.66$). Qwen3-4B exhibits strong alignment with Meta-Llama-3.1-8B (0.62), whereas Qwen3-8B shows weaker alignment with LLaMA (0.22). This mixed pattern suggests that larger models partially converge toward shared spectral partitioning of reasoning dynamics, but this consistency is not uniform across model pairs. Qwen3-1.7B behaves as an outlier among Qwen3 variants: its ARI against Qwen3-8B is near zero (-0.05), while its alignment with LLaMA matches that of Qwen3-0.6B (0.63). This pattern suggests the spectral manifold of Qwen3-1.7B



(a) (b)

Figure 3: Sensitivity analysis of W_2 inter-to-intra separation ratio. (a) Layer sensitivity: separation signal is present across all measured layers; intermediate layers yield higher ratios, while the final layer consistently produces lower values. (b) Trajectory-prefix-length sensitivity: W_2 separation ratio as a function of prefix length T. The separation signal generally strengthens as the reasoning prefix grows, reaching its maximum for the full-length reasoning trajectory T_{full}.

retains characteristics of small-model dynamics rather than adopting the partition observed in larger Qwen3 models. We leave further investigation of this non-monotonic scaling behaviour to future work.

5.5 Layer Sensitivity

We assess whether spectral separation depends on the extraction layer by recomputing the variable-rank W_2 ratio using hidden states from four depth checkpoints: $L/4$, $L/2$, $3L/4$, and the final layer L. All conditions use the same variable-rank pipeline as the main result.

Figure 3 (a) show that W_2 separation is present at every tested layer. Early and middle layers ($L/4$–$3L/4$) yield ratios of 2.06–2.67×, comparable to the main result (2.37–2.67×), demonstrating that spectral structure is not confined to the final representation. The final layer L consistently yields the lowest ratios (1.55–1.68×), suggesting that the logic-type geometry is most distinctly encoded in intermediate layers rather than the output projection. The key finding is that SIH is robust to layer choice: the spectral fingerprint captures logic-type structure throughout the model depth.

5.6 Trajectory Length Sensitivity

We evaluate whether the number of reasoning steps T used to form the trajectory matrix affects the W_2 ratio. We truncate each trace to the first $T \in \{4, 6, 8, 10\}$ steps and compare against the full trace T_{full}. All conditions use the same variable-rank pipeline.

Figure 3 (b) show that spectral separation is detectable even from the first four reasoning steps ($T = 4$), with ratios of 1.89–2.06× across models. This is a strong result: the logic-type fingerprint begins to crystallize in the earliest steps of the reasoning trace, before the model has produced a full derivation.

The full-trace ratios (T_{full}: 2.37–2.67×) match the main result, confirming consistency. Shorter prefixes yield lower but still significant ratios, indicating that the complete trajectory provides the richest spectral signal. The intermediate non-monotonicity (e.g., $T = 6$ is a local dip for Qwen3-0.6B) reflects transitional dynamics in the trajectory: as T grows from a very short prefix, the trajectory alternates between capturing coherent reasoning dynamics and transient noise before settling into the stable attractor structure visible at T_{full}.

5.7 Effective Rank Sensitivity

Here we isolate the effect of the DMD rank-cap by sweeping $r_{\max} \in \{4, 8, 16, 32\}$ on Qwen3-4B. For each $r_{\max}$, we run variable-rank DMD with a global upper bound on the number of retained

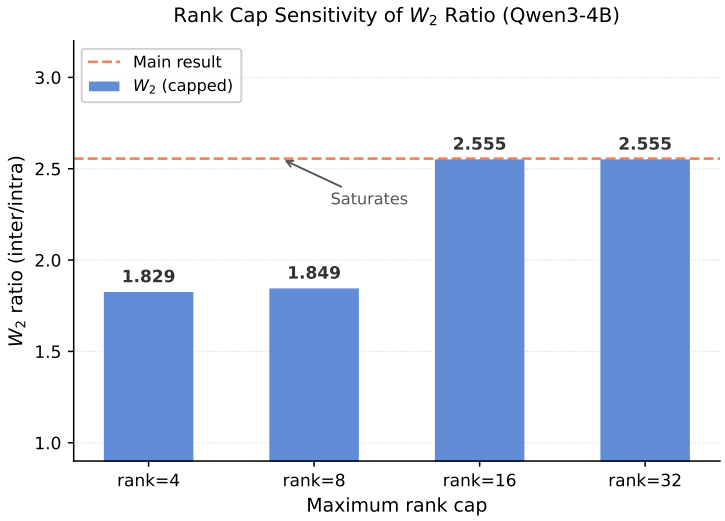


Figure 4: Rank-cap sensitivity ablation (Qwen3-4B). Inter-to-intra W_2 ratio as we sweep the global DMD rank upper bound $r_{\max} \in \{4, 8, 16, 32\}$. All conditions reuse the main-experiment sampling scheme. The ratio rises with $r_{\max}$ and saturates for $r_{\max} \geq 16$, matching the unconstrained variable-rank baseline. Saturation demonstrates that reasoning-trajectory effective ranks are bounded below 16. Tighter rank limits $(4, 8)$ discard meaningful DMD modes and artificially reduce the observed separation ratio.

singular directions, and then compute inter-to-intra W_2 following exactly the main-experiment sampling setup.

Figure 4 shows the inter-to-intra ratio rises monotonically: $r_{\max} = 4$ yields 1.829, $r_{\max} = 8$ yields 1.849, and the metric saturates starting at $r_{\max} = 16$. Increasing the cap further to $r_{\max} = 32$ yields nearly identical results, matching the unconstrained variable-rank baseline. This saturation indicates that the empirical effective rank of individual reasoning trajectories rarely exceeds 16.

Crucially, the unconstrained variable-rank result reproduces the saturated values obtained at $r_{\max} = 16$ and $r_{\max} = 32$. Once the rank cap meets or exceeds the natural effective rank of each trajectory, the capped-max-rank pipeline and adaptive per-trajectory variable-rank pipeline recover equivalent sets of DMD eigenvalues. This confirms that our variable-rank design from Equation equation 7 is not an artifact of aggressive over-truncation. The saturated plateau observed at $r_{\max} \geq 16$ *is* the variable-rank result; the lower ratio observed under tighter caps ($r_{\max} \in \{4, 8\}$) arises because those configurations discard genuine, information-bearing DMD spectral modes.

6 Conclusion

Taking a dynamical-systems perspective grounded in Koopman-operator theory, this work studies the intrinsic geometry of multi-step logical-reasoning trajectories within large-language-model latent space. Departing from conventional static-layer probing, we apply exact Dynamic Mode Decomposition to cumulative hidden-state sequences to obtain complex-valued spectral fingerprints of internal reasoning dynamics. To compare variable-cardinality DMD eigenvalue sets without zero-padding-induced distortions, we introduce a variable-rank W_2 optimal-transport metric. We formulate and empirically validate the Spectral Invariant Hypothesis, demonstrating that trajectories sharing the same logical-reasoning type exhibit tighter spectral clustering than trajectories of distinct logic types across multiple open-weight models. Spectral clustering further reveals non-monotonic scaling behaviour within model families, where the manifold structure of reasoning dynamics undergoes discrete qualitative changes as model capacity grows. Overall, our probe-free dynamical analysis establishes a complementary geometric viewpoint for understanding how logical inference manifests in LLM internal computations.

AI USE STATEMENT

We clarify that large-language models were used as auxiliary tools, exclusively for polishing and tightening the manuscript text to improve clarity and conciseness.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Science China Information Sciences*, 69(6):161101, 2026.

Jesseba Fernando and Grigori Guitchounts. Transformer dynamics: A neuroscientific approach to interpretability of large language models. *arXiv preprint arXiv:2502.12131*, 2025.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Jiaxun Gao, Him Wai Michael Ng, and Z Jane Wang. Once-more: continuous self-correction for large language models via perplexity-guided intervention. In *International Conference on Learning Representations*, volume 2026, pp. 8797–8825, 2026.

Jonas Geiping, Sean McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *Advances in Neural Information Processing Systems*, 38: 41340–41391, 2026.

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

Chaoya Jiang, Yongrui Heng, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Vlm-r^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38:63841–63869, 2026a.

Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. Tracing representation progression: Analyzing and enhancing layer-wise similarity. In *International Conference on Learning Representations*, volume 2025, pp. 1118–1143, 2025.

Xinyan Jiang, Ninghao Liu, Di Wang, and Lijie Hu. Beyond scalars: Evaluating and understanding LLM reasoning via geometric progress and stability. In *Forty-third International Conference on Machine Learning*, 2026b.

Abhinav Joshi, Divyanshu Bhatt, and Ashutosh Modi. Geometry of decision making in language models. *Advances in Neural Information Processing Systems*, 38:144139–144196, 2026.

Leonid V Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422, 1960.

Takuya Kataiwa, Cho Hakaze, and Tetsushi Ohki. Measuring intrinsic dimension of token embeddings. *arXiv preprint arXiv:2503.02142*, 2025.

Hassane Kissane, Achim Schilling, and Patrick Krauss. Probing internal representations of multiword verbs in large language models. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pp. 7–13, 2025.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Bernard O Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Tim Lawson, Lucy Farnik, Conor Houghton, and Laurence Aitchison. Residual stream analysis with multi-layer saes. In *International Conference on Learning Representations*, volume 2025, pp. 33464–33508, 2025.

Bo Li, Guanzhi Deng, Ronghao Chen, Junrong Yue, Shuo Zhang, Qinghua Zhao, Linqi Song, and Lijie Wen. REMA: A unified reasoning manifold framework for interpreting large language model, 2026.

Ryan Lucas, Kayhan Behdin, Zhipeng Wang, Qingquan Song, Rahul Mazumder, et al. Reasoning models can be accurately pruned via chain-of-thought reconstruction. In *International Conference on Learning Representations*, volume 2026, pp. 142840–142852, 2026.

Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17249–17260. IEEE, 2025.

Zeyu Ma, Wanying Wang, Guchu Zou, Mingang Chen, and Jianhong Wu. Kdp: Simplifying representation dynamics in kernel space. In *International Conference on Learning Representations*, volume 2026, pp. 155908–155932, 2026.

Igor Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*, 2025.

Phuong Minh Nguyen, Tien Dang, and Naoya Inoue. Non-interactive symbolic-aided chain-of-thought for logical reasoning. In *Proceedings of the 39th Pacific Asia Conference on Language, Information and Computation*, pp. 329–340, 2025.

Dario Picozzi. The information geometry of large language models is shared, learned, and controllable. *arXiv preprint*, 2026. doi: 10.48550/arxiv.2609.11063.

Leonardo Ranaldi, Marco Valentino, and Andre Freitas. Improving chain-of-thought reasoning via quasi-symbolic abstractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17222–17240, 2025.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, volume 2025, pp. 14855–14881, 2025.

Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of fluid mechanics*, 656:5–28, 2010.

Giovanni Servedio, Alessandro De Bellis, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. Are the hidden states hiding something? testing the limits of factuality-encoding capabilities in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6089–6104, 2025.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019.

Jonathan H Tu. *Dynamic mode decomposition: Theory and applications*. Princeton University, 2013.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Karthik Viswanathan, Yuri Gardinazzi, Giada Panerai, Alberto Cazzaniga, and Matteo Biagetti. The geometry of tokens in internal representations of large language models. *arXiv preprint arXiv:2501.10573*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Zhaomin Wu, Mingzhe Du, See-Kiong Ng, and Bingsheng He. Beyond prompt-induced lies: Investigating llm deception on benign prompts. In *International Conference on Learning Representations*, volume 2026, pp. 88417–88445, 2026.

Menglin Yang, Aosong Feng, Bo Xiong, Jiahong Liu, Irwin King, Rex Ying, et al. Hyperbolic fine-tuning for large language models. *Advances in Neural Information Processing Systems*, 38: 29613–29648, 2026.

Hengyuan Zhang, Chenming Shang, Sizhe Wang, Dongdong Zhang, Yiyao Yu, Feng Yao, Renliang Sun, Yujiu Yang, and Furu Wei. Shifcon: Enhancing non-dominant language capabilities with a shift-based multilingual contrastive framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4818–4841, 2025.

Yu Zheng, Yuan Yuan, Yue Zhuo, Yong Li, and Paolo Santi. Probing neural topology of large language models, 2026.

Yufa Zhou, Yixiao Wang, Xunjian Yin, Shuyan Zhou, and Anru Zhang. The geometry of reasoning: Flowing logics in representation space. In *International Conference on Learning Representations*, volume 2026, pp. 72711–72739, 2026.

APPENDIX

A ZERO-PADDING ABLATION

As described in Section 3.3, we compare two strategies for handling variable-rank spectral fingerprints:

1. **Variable-rank** (ours): truncate both spectra to $k = \min(r_a, r_b)$ eigenvalues sorted by $|\lambda|$ descending, then apply Hungarian matching.
2. **Zero-padded**: extend the shorter spectrum to the longer rank $k = \max(r_a, r_b)$ by appending zeros, then apply Hungarian matching over all k elements.

Table 4 reports the inter/intra W_2 ratio for each strategy.

Table 4: **Variable-rank vs. zero-padded W_2 ratios.** At layer $L/2$, effective ranks are compact and similar across instances, so zero-padding adds few structural zeros and yields ratios modestly above the variable-rank baseline.

Model	Variable-rank ratio	Zero-padded ratio
Qwen3-0.6B	2.49	2.61
Qwen3-1.7B	2.53	2.64
Qwen3-4B	2.56	2.67
Qwen3-8B	2.37	2.51
Meta-Llama-3.1-8B	2.67	2.77

The zero-padded ratios (2.51–2.77×) are modestly higher than the variable-rank ratios (2.37–2.67×) for every model, with a lift of 0.10–0.14 ratio points. At layer $L/2$, effective ranks are small and similar across instances, so the two spectra in each pair are nearly the same length; padding adds at most a few structural zeros. The additional matched pairs (small genuine eigenvalues against structural zeros) contribute positively to inter-class discriminability, yielding a modest lift rather than suppression. We adopt variable-rank matching as the primary metric because it avoids introducing structural zeros and provides a more conservative, interpretable lower bound on spectral separation.

Table 5: Example instance from the Reasoning-Flow dataset. One abstract logical template is instantiated into multiple topics and languages, preserving identical logical inference steps while varying surface-level linguistic content.

Abstract symbolic template	**Natural language (English, software topic)**	**Natural language (German, software topic)**
[1] $A \to B$	[1] If the project contains a valid package manifest, then dependencies install successfully.	[1] Wenn das Projekt eine gültige Paketmanifestdatei enthält, dann lassen sich die Abhängigkeiten erfolgreich installieren.
[2] $B \to C$	[2] If dependencies install successfully, then the application builds successfully.	[2] Wenn die Abhängigkeiten erfolgreich installiert sind, dann wird die Anwendung erfolgreich gebaut.
[3] $\forall x\,(H(x) \to J(x))$	[3] For every module x, if x has tests that pass, then CI marks x as tested.	[3] Für jedes Modul x gilt: Wenn x bestandene Tests hat, dann markiert die CI x als getestet.
[4] $H(a)$	[4] The auth module has tests that pass.	[4] Das Auth-Modul hat bestandene Tests.
[5] A	[5] The project contains a valid package manifest.	[5] Das Projekt enthält eine gültige Paketmanifestdatei.
[6] B (from [1] and [5])	[6] Therefore, dependencies install successfully.	[6] Daher lassen sich die Abhängigkeiten erfolgreich installieren.
[7] C (from [2] and [6])	[7] Therefore, the application builds successfully.	[7] Daher wird die Anwendung erfolgreich gebaut.
[8] $J(a)$ (from [3] and [4])	[8] Therefore, CI marks the auth module as tested.	[8] Daher markiert die CI das Auth-Modul als getestet.
[9] $C \wedge J(a)$ (from [7] and [8])	[9] Therefore, the application builds successfully and CI marks the auth module as tested.	[9] Daher wird die Anwendung erfolgreich gebaut und die CI markiert das Auth-Modul als getestet.

B Reasoning-Flow Dataset Details

Reasoning-Flow (Zhou et al., 2026) is a publicly released benchmark dataset constructed to examine whether LLM reasoning trajectories are governed by underlying logical structure rather than superficial semantic content. The dataset contains 30 distinct logical structures. Each logical template is instantiated across 20 topical domains and rendered in four languages (English, Chinese, German, Japanese), yielding a total corpus of 2,430 reasoning sequences. Each reasoning instance produces a trajectory with 8 to 16 discrete reasoning steps. Table 5 shows an example instance from the dataset.

C Permutation Sensitivity Analysis

To assess whether the permutation test result depends on the number of instances sampled per logic class, we re-ran the test on Qwen3-4B with $n_{\text{sample}} \in \{10, 15, 20\}$. For each value, the observed ratio T_{obs} is computed with the same intra aggregation as the main result (Section 5).

Table 6: **Permutation test stability across sample sizes (Qwen3-4B).** The observed inter/intra ratio is stable across a $2\times$ range of sample sizes; $p = 0/500$ at all three settings.

n_{sample}	Instances	Observed ratio	p-value
10	300	2.601	$0/500$
15	450	2.555	$0/500$
20	600	2.525	$0/500$

As shown in Table 6, the observed ratio varies by less than 0.08 across the three settings (2.53–2.60$\times$), confirming that the signal is stable and not an artifact of a particular sample size. The $n_{\text{sample}} = 15$ row reproduces the main result (2.555$\times$), serving as an anchor point. In all three cases, zero of 500 permutations achieve the observed ratio, establishing $p < 0.002$ regardless of sample size.

D Mathematical Propositions and Proofs

This appendix provides formal proofs for the propositions stated in the main text, alongside additional theoretical properties characterizing our Exact DMD application and the variable-rank Wasserstein distance.

D.1 Proof of Proposition 1 (Comparison of Variable-Rank and Zero-Padded Distances)

Proof. By definition, the variable-rank strategy computes the optimal alignment over the top-k eigenvalues. Let $\pi^* : [k] \to [k]$ be the optimal bijection that achieves the minimum in the variable-rank distance:

$$k\big(W_2^{\mathrm{var}}\big)^2 = \sum_{i=1}^{k} \big|\lambda_i^{(a)} - \lambda_{\pi^*(i)}^{(b)}\big|^2. \tag{9}$$

The zero-padded strategy extends the shorter spectrum to length K. Let the padded spectrum for $\mathcal{F}_b$ be defined as $\lambda_j^{(b,\mathrm{pad})} = \lambda_j^{(b)}$ for $j \le k$, and $\lambda_j^{(b,\mathrm{pad})} = 0$ for $k < j \le K$. The zero-padded distance applies Hungarian matching over all K elements, minimizing the cost over all bijections $\sigma : [K] \to [K]$:

$$K\big(W_2^{\mathrm{pad}}\big)^2 = \min_{\sigma} \sum_{i=1}^{K} \big|\lambda_i^{(a)} - \lambda_{\sigma(i)}^{(b,\mathrm{pad})}\big|^2. \tag{10}$$

We construct a specific, valid bijection $\hat{\sigma}$ over K elements by extending π^* such that it exactly mimics the variable-rank matching for the first k elements, while assigning the remaining $K - k$ elements to the appended zeros. Specifically:

$$\hat{\sigma}(i) = \begin{cases} \pi^*(i) & \text{if } i \le k, \\ i & \text{if } i > k. \end{cases} \tag{11}$$

The transport cost associated with this specific bijection $\hat{\sigma}$ evaluates to:

$$C(\hat{\sigma}) = \sum_{i=1}^{k} \big|\lambda_i^{(a)} - \lambda_{\hat{\sigma}(i)}^{(b,\mathrm{pad})}\big|^2 + \sum_{i=k+1}^{K} \big|\lambda_i^{(a)} - \lambda_{\hat{\sigma}(i)}^{(b,\mathrm{pad})}\big|^2 \tag{12}$$

$$= \sum_{i=1}^{k} \big|\lambda_i^{(a)} - \lambda_{\pi^*(i)}^{(b)}\big|^2 + \sum_{i=k+1}^{K} \big|\lambda_i^{(a)} - 0\big|^2 \tag{13}$$

$$= k\big(W_2^{\mathrm{var}}\big)^2 + \sum_{i=k+1}^{K} \big|\lambda_i^{(a)}\big|^2. \tag{14}$$

Because the Hungarian algorithm finds the absolute minimum over all possible matching permutations, the optimal zero-padded cost cannot exceed the cost of our constructed bijection $\hat{\sigma}$. Therefore:

$$K\big(W_2^{\mathrm{pad}}\big)^2 \le C(\hat{\sigma}) = k\big(W_2^{\mathrm{var}}\big)^2 + \sum_{i=k+1}^{K} \big|\lambda_i^{(a)}\big|^2. \tag{15}$$

This mathematically demonstrates that the zero-padded method is functionally equivalent to the variable-rank distance plus a structural penalty term. This penalty term artificially inflates the distance by absorbing the magnitudes of the tail eigenvalues of the higher-rank trajectory, confounding pure dynamical misalignment with differences in intrinsic effective rank. □

D.2 Topological Properties of the Variable-Rank W_2

In practice, spectral fingerprints extracted from LLM reasoning trajectories exhibit heterogeneous effective ranks. This proposition clarifies the topological properties of our variable-rank W_2 functional: it forms a full metric when comparing fingerprints of identical rank, and remains a valid

pairwise comparison tool under rank heterogeneity. Even though truncation breaks the triangle inequality for mixed-rank inputs, this property is a deliberate trade-off: it enables us to focus comparison exclusively on dominant dynamical modes and avoid distortions introduced by zero-padding. This theoretical characterization supports our experimental design, where we aggregate pairwise distances into inter-class / intra-class mean ratios rather than relying on metric-dependent downstream algorithms.

Proposition 2 (Metric Properties of Truncated Wasserstein)**.** *Let $d(\alpha_a, \alpha_b) = W_2(\mathcal{F}(\alpha_a), \mathcal{F}(\alpha_b))$ as defined in Eq. (7). For trajectories with identical effective ranks ($r_a = r_b$), d satisfies all properties of a metric space (symmetry, non-negativity, identity of indiscernibles, and the triangle inequality). However, for trajectories with variable effective ranks ($r_a \neq r_b$), d mathematically acts as a* semimetric *because the truncation step inherently violates the triangle inequality.*

Proof. Non-negativity ($d \geq 0$) and symmetry ($d(\alpha_a, \alpha_b) = d(\alpha_b, \alpha_a)$) follow trivially from the symmetric formulation of $k = \min(r_a, r_b)$ and the underlying Euclidean distance in $\mathbb{C}$.

Case 1 ($r_a = r_b = r$): Here, $k = r$. No eigenvalue truncation occurs. Equation (7) simplifies exactly to the standard 2-Wasserstein distance between two discrete empirical probability measures with equal uniform mass. Because the standard 2-Wasserstein distance is a well-established metric on the space of probability measures with finite second moments, the triangle inequality holds.

Case 2 ($r_a \neq r_b$): We prove the violation of the triangle inequality by constructing a counterexample. We wish to find three spectral fingerprints $\mathcal{F}_A, \mathcal{F}_B, \mathcal{F}_C$ such that $d(\mathcal{F}_A, \mathcal{F}_C) > d(\mathcal{F}_A, \mathcal{F}_B) + d(\mathcal{F}_B, \mathcal{F}_C)$. Consider three fingerprints sorted by magnitude:

- $\mathcal{F}_A = \{10, 5\}$ (Rank $r_A = 2$)
- $\mathcal{F}_B = \{10\}$ (Rank $r_B = 1$)
- $\mathcal{F}_C = \{10, -5\}$ (Rank $r_C = 2$)

We compute the pairwise variable-rank distances:

1. $d(\mathcal{F}_A, \mathcal{F}_B)$: $k = \min(2, 1) = 1$. We truncate $\mathcal{F}_A$ to $\{10\}$.
   $d(\mathcal{F}_A, \mathcal{F}_B) = \sqrt{|10 - 10|^2/1} = 0$.
2. $d(\mathcal{F}_B, \mathcal{F}_C)$: $k = \min(1, 2) = 1$. We truncate $\mathcal{F}_C$ to $\{10\}$.
   $d(\mathcal{F}_B, \mathcal{F}_C) = \sqrt{|10 - 10|^2/1} = 0$.
3. $d(\mathcal{F}_A, \mathcal{F}_C)$: $k = \min(2, 2) = 2$. No truncation occurs.
   $d(\mathcal{F}_A, \mathcal{F}_C) = \sqrt{\frac{1}{2}\left(|10 - 10|^2 + |5 - (-5)|^2\right)} = \sqrt{\frac{1}{2}(0 + 100)} = \sqrt{50} \approx 7.07$.

Here, $d(\mathcal{F}_A, \mathcal{F}_B) + d(\mathcal{F}_B, \mathcal{F}_C) = 0 + 0 = 0$. However, $d(\mathcal{F}_A, \mathcal{F}_C) = 7.07$. Since $7.07 \not\leq 0$, the triangle inequality is fundamentally violated.

This proves that while variable-rank W_2 effectively identifies similarities among dominant dynamical modes, the truncation step inherently creates low-rank "bottlenecks" (such as $\mathcal{F}_B$) that short-circuit paths between higher-rank representations, legally stripping it of formal metric status in the variable-rank regime. □

D.3 Orthogonal Invariance

A crucial property of the DMD-derived spectral fingerprint $\mathcal{F}(\alpha)$ is its invariance to isometric transformations of the LLM's representation space. This proposition guarantees that our spectral characterization isolates the intrinsic temporal dynamics of the reasoning process, making it completely independent of the arbitrary basis chosen by the LLM's layers. We formalize this as follows:

Proposition 3 (Orthogonal Invariance of the Spectral Fingerprint)**.** *Let $Q \in \mathbb{R}^{d \times d}$ be an arbitrary orthogonal matrix (i.e., $Q^\top Q = I$). If the hidden-state trajectory Y is transformed via $Y' = QY$ (representing a global rotation or reflection of the latent space), the resulting spectral fingerprint is strictly preserved: $\mathcal{F}'(\alpha) = \mathcal{F}(\alpha)$.*

Proof. Let the rotated trajectory matrix be $Y' = QY$. Consequently, the snapshot matrices are transformed as $Y'_1 = QY_1$ and $Y'_2 = QY_2$. Consider the economy SVD of the original snapshot matrix: $Y_1 = U\Sigma V^\top$. The SVD of the rotated snapshot matrix becomes:

$$Y'_1 = (QU)\Sigma V^\top = U'\Sigma V^\top, \tag{16}$$

where $U' = QU$. Because Q is orthogonal, U' retains orthonormality, since $(U')^\top U' = U^\top Q^\top QU = U^\top IU = I$. The singular values Σ and right singular vectors V are completely unaffected.

Following Eq. (4), the companion matrix for the transformed system is computed as:

$$\tilde{A}' = (U')^\top Y'_2 V\Sigma^{-1} = (QU)^\top (QY_2)V\Sigma^{-1} = U^\top (Q^\top Q)Y_2 V\Sigma^{-1}. \tag{17}$$

Since $Q^\top Q = I$, this simplifies to:

$$\tilde{A}' = U^\top Y_2 V\Sigma^{-1} = \tilde{A}. \tag{18}$$

Because the companion matrices are identical ($\tilde{A}' = \tilde{A}$), their eigenvalue spectra are naturally identical. Thus, $\mathcal{F}'(\alpha) = \mathcal{F}(\alpha)$. □