Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Task-Dependent Intrinsic Dimensionality Across Transformer Layers: An Exploratory Study of Qwen2.5-1.5B

Mitra Varun Gogulapati

mitravarun22@gmail.com

Kakatiya University

Research Article

Keywords: representation geometry; intrinsic dimension; participation ratio; transformers; Qwen; exploratory analysis.

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11703266/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Task-Dependent Intrinsic Dimensionality Across Transformer Layers: An Exploratory Study of Qwen2.5-1.5B

Mitra Varun Gogulapati
Independent Researcher
Hyderabad, India
mitravarun2702@gmail.com

October 2026

Abstract

We ask whether task categories differ in the dimensionality of a pretrained language model's internal representations and whether their relationship changes with depth. Last-input-token representations are extracted from Qwen2.5-1.5B for 200 prompts spanning factual recall, arithmetic, reasoning, and coding. Representation geometry is summarized using participation ratio, covariance-spectrum entropy effective rank, principal-component variance thresholds, and the TWO-NN local intrinsic-dimension estimator. The primary held-out analysis contains 12 prompts from two template families per category. For raw representations, the reasoning-minus-recall participation-ratio contrast is positive at blocks 1–20 and negative at blocks 21–28. At block 24, participation ratio is 2.869 for factual recall and 1.397 for reasoning; after the final block and normalization, it is 2.235 and 2.126, respectively. The observations establish a descriptive depth-dependent relationship in these prompt sets, rather than a uniform ordering of task categories. Limited held-out family counts prevent reliable primary family-bootstrap intervals, and insufficient length overlap prevents a joint matched-length comparison. We report a bounded exploratory experiment with its analysis code and limitations.

Keywords: representation geometry; intrinsic dimension; participation ratio; transformers; Qwen; exploratory analysis.

1 Introduction

A transformer representation can have thousands of coordinates while most variation across inputs lies in a much smaller set of directions. Ambient width therefore does not specify the effective dimensionality of an observed collection of hidden states. Its covariance spectrum and local neighborhood geometry describe different aspects of how input-dependent representations are organized.

Prior research has examined intrinsic dimension across neural-network layers and identified depth-dependent geometric structure in language transformers (Ansuini et al., 2019; Cheng et al., 2025). This motivates a narrower question: within one fixed model, do prompt sets assigned to different task categories exhibit different dimensionality profiles? We compare factual recall, arithmetic, reasoning, and coding without changing model weights.

Three questions guide the analysis. First, how do dimensionality summaries vary across depth? Second, how does the reasoning–recall contrast change across layers? Third, how sensitive are the observations to estimator choice, representation preprocessing, sample size, and

template composition? The last question also requires reporting controls that the dataset cannot support.

The contributions are a reproducible analysis workflow, a layer-wise comparison of four operational task categories, and an analysis of the limits of a small experiment. No new dimensionality estimator is proposed. No answers are generated, and our experiment does not evaluate answer correctness, model training, or causal reasoning mechanisms. The term "task-dependent" in the title refers to observed differences between the constructed prompt sets; we cannot identify a causal task effect from this descriptive comparison.

2 Related work

Representation dimensionality. Ansuini et al. (2019) studied the intrinsic dimension of representations in deep neural networks. Cheng et al. (2025) identified a high-dimensional intermediate phase across several language transformers and input datasets. Their broader study motivates measuring depth profiles, but does not imply that a task contrast in the present small model has the same explanation.

Contextual geometry. Ethayarajh (2019) showed that contextualized representations in several language models are anisotropic. Such results motivate attention to variance concentration: a high-dimensional vector need not yield a high participation ratio. Centered covariance spectra in our experiment measure variation around the sample mean, rather than directly measuring the angular anisotropy of uncentered embeddings.

Different objects called intrinsic dimension. Aghajanyan et al. (2021) studied low-dimensional parameterizations of language-model fine-tuning. That object lies in parameter space. Here the weights remain fixed and the object is a finite cloud of activation vectors. These dimensions should not be conflated.

Estimators. Participation ratio summarizes spectral concentration and has been used in neural population analysis (Mazzucato et al., 2016). Entropy effective rank follows an entropy-based construction (Roy and Vetterli, 2007), with a stated covariance-eigenvalue convention here. Principal component thresholds summarize explained variance (Jolliffe and Cadima, 2016). TWO-NN instead uses nearest-neighbor distance ratios (Facco et al., 2017). Agreement is not guaranteed because the estimators describe different geometric properties.

3 Dataset and scope

3.1 Task operationalizations

The dataset contains 200 prompts, with 50 per category. Factual recall uses chemical-element fixtures, including symbols, atomic numbers, and proton counts. Arithmetic uses explicit calculations with computed answer fixtures. Reasoning uses Boolean or relational problems whose fixtures follow the stated premises. Coding asks for the values returned by short Python expressions. Consequently, "coding" means Python output prediction in this experiment, not open-ended program synthesis, and "factual recall" is a narrow chemistry domain.

Prompts are generated from eight named template families per category. A family fixes a syntactic frame while varying entities, strings, numbers, or premises. Every prompt requests a short answer and shares a final answer-request suffix. The expected answers are dataset

fixtures; they are not measured model outputs. Appendix A provides one actual example from every family, and the source archive includes all 200 records.

3.2 Partition and sample counts

Families 0–5 form the exploratory partition and families 6–7 form the held-out evaluation partition. This produces 38 exploratory and 12 held-out prompts per category (Table 1). The primary spectral and local-dimension summaries use the held-out partition. Thus 200 prompts were extracted, but the primary analysis contains 48 vectors per layer, separated into four 12-point clouds. The partitions do not train or fine-tune the model, and the held-out designation does not imply preregistered confirmatory inference.

Category	Total	Exploratory	Held-out	Held-out families
Factual recall	50	38	12	2
Arithmetic	50	38	12	2
Reasoning	50	38	12	2
Coding	50	38	12	2

Table 1: Dataset composition. Each category has eight families overall: six exploratory and two held-out. No model training is performed.

3.3 Sample size

We used 200 prompts to keep extraction and checking manageable. The sample size is not power-based.

4 Model and activation extraction

The experiment uses the pretrained base model `Qwen/Qwen2.5-1.5B` (Qwen Team, 2024), with 28 transformer blocks and hidden width 1,536. The resolved revision is `8faed761d45a2633 40a0528343f099c05c9a4323`. Extraction ran in a hosted GPU runtime on a Tesla T4 GPU, using float32, batch size one, eager attention, and seed 42. The recorded software environment is given in Appendix G.

For prompt i, let $m_{it} \in \{0, 1\}$ be its attention mask and let $z_{it}^{(\ell)}$ be the returned hidden vector at token position t and output index ℓ. The recorded representation is

$$t_i^* = \max\{t : m_{it} = 1\}, \qquad h_i^{(\ell)} = z_{it_i^*}^{(\ell)} \in \mathbb{R}^{1536}. \tag{1}$$

Only the input prompt is processed. No generated continuation is included. Returned index 0 is the embedding output, indices 1–27 correspond to block outputs, and index 28 includes the final transformer block and final RMS normalization. The last output is therefore not a measurement of the final block before normalization.

Extraction completed for all 200 prompts at all 29 output indices; the metadata records completion. An earlier float16 attempt produced nonfinite representations. The successful float32 run is the basis of this paper. This is a run-specific numerical observation, not a claim that all half-precision uses of Qwen fail. Extraction checks shape and finiteness before advancing a checkpoint. Stored activation arrays would occupy 35,635,200 bytes (about 34 MiB), excluding model weights and transient tensors. The results archive contains metadata and derived results, but not those activation arrays.

At index 0, the last token's embedding does not depend on the preceding context. The shared suffix yields repeated identical final-token embeddings and zero-variance clouds in the outputs. Undefined dimensionality there is an expected degeneracy, not evidence of extraction failure.

5 Mathematical methods

For one category at one layer, stack N vectors in $H \in \mathbb{R}^{N \times d}$. Let $\bar{h} = N^{-1} \sum_i h_i$, and center the rows:

$$X = H - \mathbf{1}\bar{h}^\top, \qquad C = \frac{X^\top X}{N-1}. \tag{2}$$

An economy singular value decomposition in float64 supplies singular values s_i and covariance eigenvalues $\lambda_i = s_i^2/(N-1)$. The nonzero spectrum of the much larger feature covariance can therefore be obtained without allocating a $d \times d$ matrix.

5.1 Spectral measures

Participation ratio (PR), the primary summary, is

$$D_{\mathrm{PR}} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}. \tag{3}$$

For positive total variance, it equals one when a single eigenvalue carries all variance and equals r when r nonzero eigenvalues are equal. It measures variance concentration, not a count of reasoning operations.

The covariance-spectrum entropy effective rank is

$$p_i = \frac{\lambda_i}{\sum_j \lambda_j}, \qquad D_{\mathrm{ER}} = \exp\left(- \sum_{i:p_i>0} p_i \log p_i \right). \tag{4}$$

The entropy construction is related to Roy and Vetterli (2007); their original normalization uses singular values, whereas this paper explicitly normalizes covariance eigenvalues. The two conventions need not agree.

For a variance target $q \in \{0.80, 0.90, 0.95, 0.99\}$, the PCA component threshold is

$$D_q = \min\left\{ k : \frac{\sum_{i=1}^k \lambda_i}{\sum_j \lambda_j} \geq q \right\}. \tag{5}$$

These are separate spectral summaries. They are undefined if total variance is zero.

For centered finite data,

$$\mathrm{rank}(X) \leq \min(N-1, d). \tag{6}$$

Thus the primary clouds with $N = 12$ have at most 11 nonzero centered directions. PR, entropy effective rank, and PCA component thresholds are bounded accordingly in exact arithmetic. A low estimate from this sample is not a claim that the full model's activation distribution has only a few dimensions.

5.2 Local dimension via TWO-NN

Let $r_1(i)$ and $r_2(i)$ be Euclidean distances from point i to its first and second neighbors, and $\mu_i = r_2(i)/r_1(i)$. Under the local assumptions of the estimator, the ratio distribution satisfies (Facco et al., 2017)

$$F(\mu) = 1 - \mu^{-D}, \qquad -\log(1 - F(\mu)) = D \log \mu. \tag{7}$$

The implementation sorts $\log \mu_i$, uses midpoint plotting positions $F_i = (i - \frac{1}{2})/N$, and omits the largest 10% of ratios. Writing the retained coordinates as $(a_i, b_i) = (\log \mu_i, -\log(1 - F_i))$, its origin-fit estimate is

$$\widehat{D}_{\text{TWO-NN}} = \frac{\sum_i a_i b_i}{\sum_i a_i^2}. \tag{8}$$

At $N = 12$, only ten fit points remain. Duplicate points invalidate ratios, fewer than ten inputs are rejected, and fit $R^2 < 0.8$ is flagged as poor. Passing this diagnostic does not establish estimator accuracy at such small sample size. TWO-NN is not constrained to equal covariance effective dimension or the finite sample rank.

5.3 Preprocessing and contrast

Raw representations are primary. Two sensitivity modes use row normalization $h_i/\|h_i\|_2$ and feature standardization $(h_{ij} - \bar{h}_j)/s_j$, respectively, with degenerate coordinates handled by the implementation. These transformations change the geometry and are reported separately.

The principal descriptive contrast at block ℓ is

$$\Delta^{(\ell)}_{\text{PR}} = D^{(\ell)}_{\text{PR,reasoning}} - D^{(\ell)}_{\text{PR,recall}}. \tag{9}$$

A positive value describes the two sampled clouds at that layer, without implying a population-level difference or causal task effect.

6 Statistical treatment

Within-family prompts share construction rules, so treating every prompt as an independent dimension observation would be inappropriate. Each dimensionality estimate is a statistic of an entire cloud. The pipeline supports resampling whole families and recomputing spectral contrasts, but requires at least four families per category for primary family-bootstrap inference. The held-out categories have only two. Consequently, primary intervals are withheld; the configured 200 bootstrap replicates do not imply that 200 valid primary replicates were executed.

Permutation tests are disabled because exchangeability of the category-specific families has not been established. No primary p-values or adjusted p-values are reported. The absence of these quantities is neither a zero p-value nor evidence of no effect. The pipeline's multiple-testing capability does not turn these descriptive observations into tested findings. Layer outputs are also dependent and cannot be treated as 28 independent experimental replications.

The exploratory partition contains six families per category, and some exploratory control curves include family-bootstrap intervals. Those intervals concern that different partition and do not supply missing primary uncertainty. TWO-NN replacement bootstrap intervals are disabled because repeated sampled points invalidate its neighbor ratios. All numerical claims below are descriptive.

7 Results

7.1 Layer-wise profiles

Figure 1 shows all four dimensionality summaries across depth. The embedding output is degenerate; the contextualized outputs show nonconstant profiles. Small PR values indicate concentration of centered variance, even though the vectors each contain 1,536 coordinates. The result should be read subject to the finite sample bound in Equation (6).

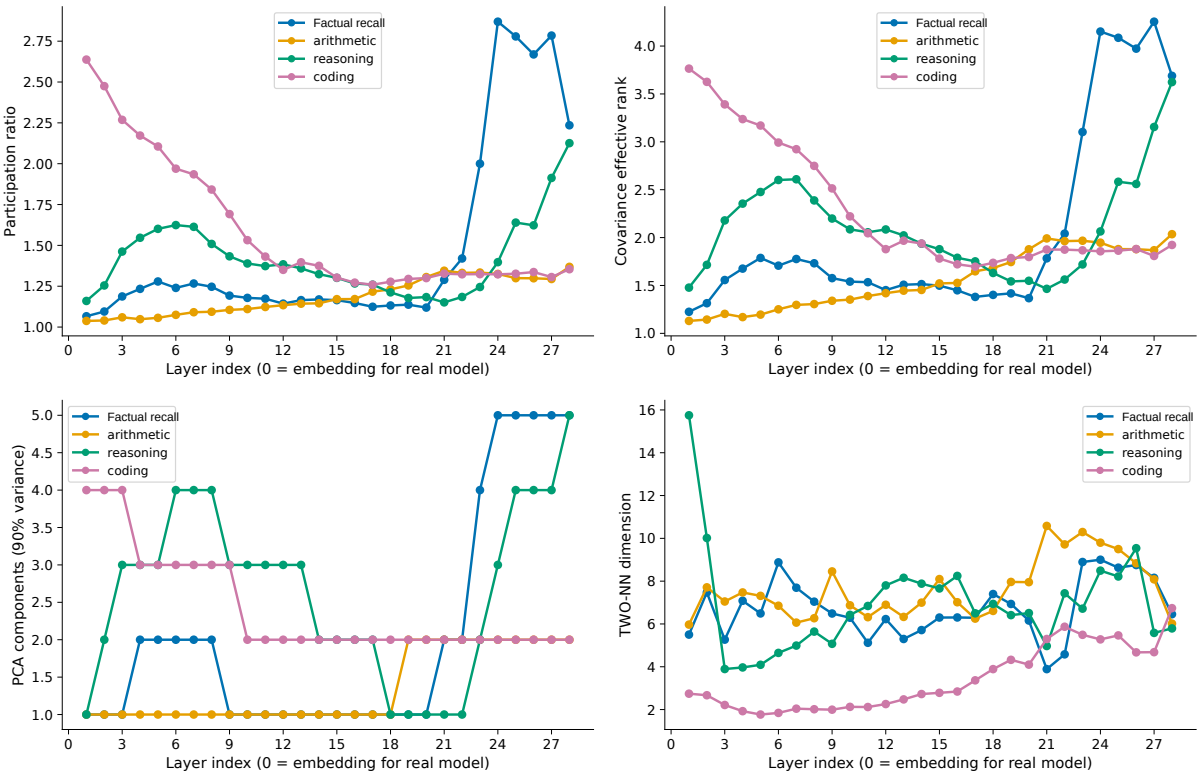


Figure 1: Layer-wise dimensionality profiles for the held-out raw representations, with 12 prompts per category. Index 0 is the embedding output; index 28 includes final RMSNorm. Missing embedding estimates are undefined. The panels describe different quantities and do not share a common notion of dimension.

7.2 Reasoning–recall contrast

The primary raw PR contrast is positive at blocks 1–20 and negative at blocks 21–28 (Figure 2). Its largest positive value is +0.384 at block 6. The most negative value is −1.472 at block 24. Table 2 provides selected values from the CSVs, including blocks adjacent to the sign crossing.

Block	Recall	Reasoning	Δ_{PR}
1	1.066	1.160	+0.094
6	1.240	1.624	+0.384
20	1.119	1.183	+0.064
21	1.289	1.151	-0.138
24	2.869	1.397	-1.472
28 + norm	2.235	2.126	-0.110

Table 2: Selected primary raw PR values, $N = 12$ per category. Values are descriptive; primary intervals and p-values are unavailable. Block selection summarizes the observed profile and is not an independent test.

At block 20, recall PR is 1.119 and reasoning PR is 1.183. At block 24, the values are 2.869 and 1.397, respectively. Thus the late negative contrast is associated with a larger increase in recall PR over those outputs. At the final returned output, recall and reasoning are closer, at 2.235 and 2.126. Final normalization is included there, so the narrowed gap cannot be assigned

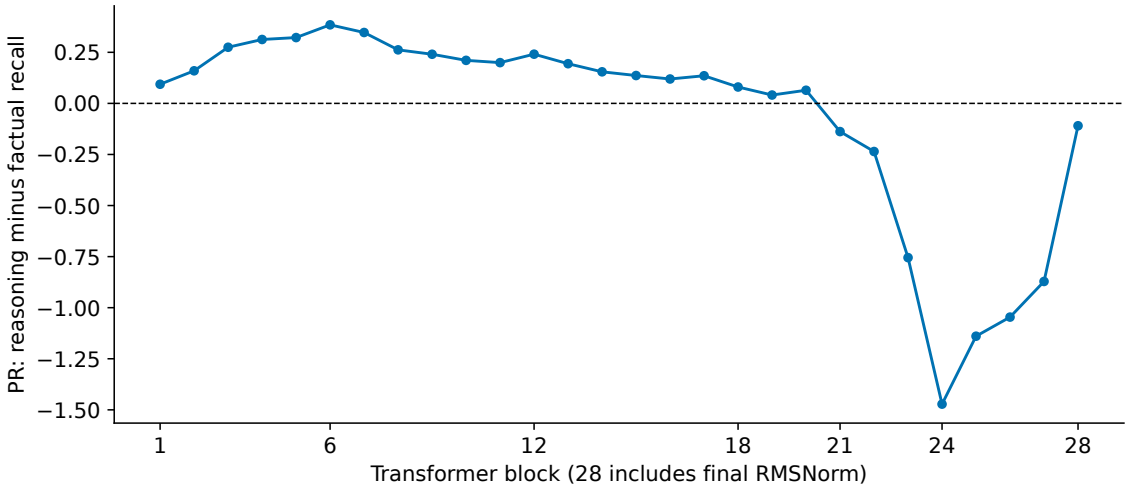


Figure 2: Reasoning-minus-recall PR contrast computed from the comparison table. The dashed line denotes zero. There are no confidence bands because primary uncertainty was unavailable. The sign crossing is descriptive, not a statistically identified phase transition.

solely to the final block's computation.

7.3 Final-output summaries and spectra

Table 3 compares all four categories at output 28. Factual recall and reasoning both require five components to reach 90% variance, whereas arithmetic and coding require two. Nevertheless, their TWO-NN values do not follow the same ordering as PR. This is an observed difference between a local neighbor-based fit and spectral concentration, not proof that either estimator is correct or incorrect.

Category	PR	Effective rank	D_{90}	D_{95}	TWO-NN
Factual recall	2.235	3.689	5	6	6.455
Arithmetic	1.369	2.036	2	5	6.017
Reasoning	2.126	3.624	5	7	5.792
Coding	1.354	1.923	2	3	6.738

Table 3: Final returned output, including RMSNorm: raw held-out estimates with $N = 12$ per category. D_{90} and D_{95} are PCA thresholds. All four displayed TWO-NN fits pass the implemented fit diagnostic, but remain small-sample estimates.

7.4 Length checks and unavailable controls

Token lengths differ by category. Across all 50 prompts per category, median lengths are 16 for recall, 17 for arithmetic, 29 for reasoning, and 31 for coding. Figure 4 shows the held-out distributions; Appendix B provides full-dataset quantiles. The configured four-token-bin selection found no bins shared across all four categories and retained no joint matched sample. Equal prompt counts therefore do not constitute a length control. The primary comparison remains potentially confounded by length and family structure.

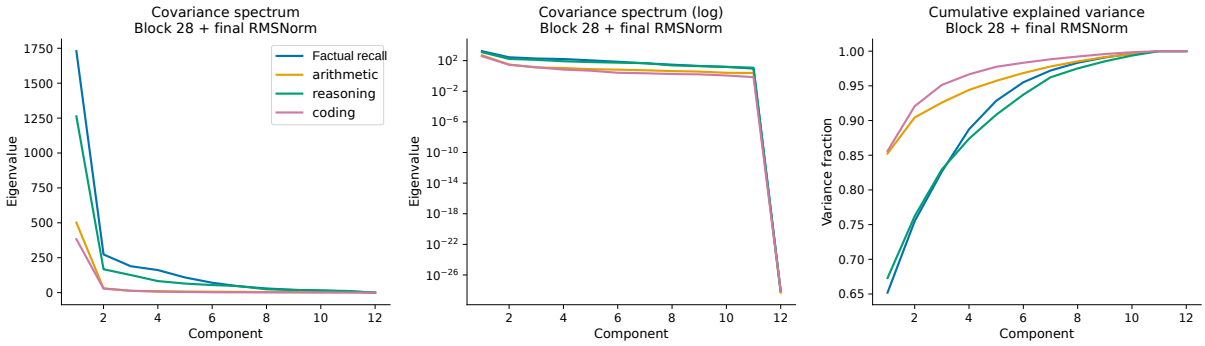


Figure 3: Covariance eigenvalue spectra and cumulative explained variance at output 28 for raw held-out clouds. The logarithmic panel exposes small tail eigenvalues, including possible numerical residuals near zero. The cumulative panel describes the PCA thresholds in Table 3; it is not a local manifold-dimension estimate.

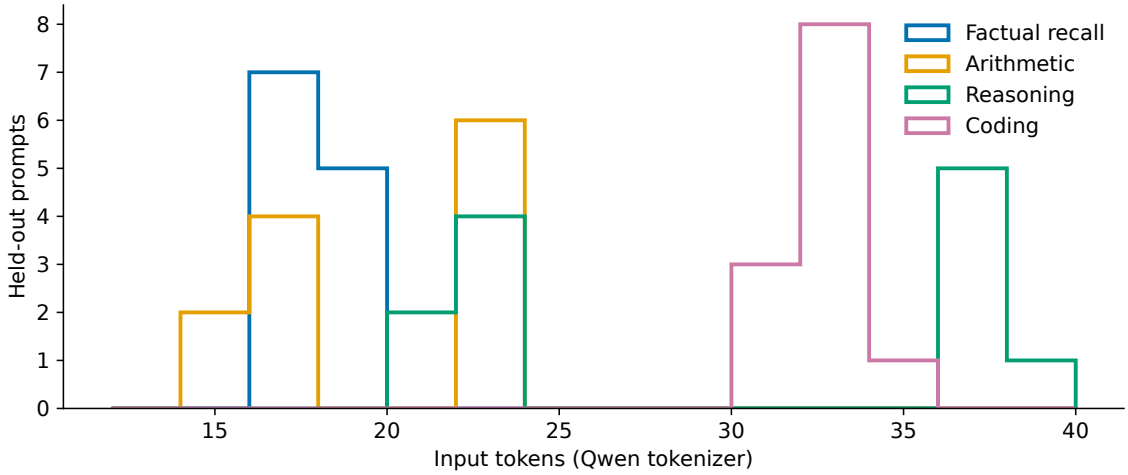


Figure 4: Input-token distributions in the actual held-out partition, 12 prompts per category, measured by the Qwen tokenizer. Histogram bins are for display; the analysis separately attempted four-token-bin matching and found no common bins across all four categories.

8 Robustness and diagnostics

Preprocessing. At block 24, reasoning-minus-recall PR is -1.472 for raw vectors, -1.363 after unit normalization, and -1.773 after feature standardization. At block 21 all three are negative, while at block 6 all are positive. Appendix E provides selected contrasts. These checks show that those particular signs are not unique to the raw scaling convention; they do not eliminate length, vocabulary, or template confounds.

Subsampling. Five repeated common-count subsamples are configured. For the 12-point held-out clouds, only the requested size of eight is available; requested sizes 16 and 32 exceed the available category size. The final-output robustness panel therefore supports only a narrow subsampling check, not a sample-size convergence curve. Its shading is a descriptive range over repeats and is not a confidence interval.

Template comparisons. The original control figure compares exploratory and evaluation partitions. Differences between those curves can reflect both family composition and sample count (38 versus 12), so they do not isolate a pure template effect. Their uncertainty cannot be transferred between partitions.

Degenerate and unstable estimates. The summary records 256 flagged estimate entries: 144 zero-variance, 76 poor-fit, and 36 duplicate-point entries. These are counts across metrics, subsets, layers, and preprocessing modes, not 256 failed prompts. A poor-fit TWO-NN value may still be present numerically but is flagged. The diagnostic categories are kept distinct from unperformed inference. Additional figures are retained in the appendix for reproducibility.

9 Discussion

The strongest supported observation is that the raw reasoning–recall PR ordering is not uniform across depth. The same small held-out prompt sets yield a positive contrast in early and middle blocks and a negative contrast in the later blocks. A statement that reasoning is universally higher-dimensional is therefore inconsistent with this particular observed profile.

A geometric explanation is that different categories distribute centered variance differently as the model transforms the prompts. However, category labels also track prompt construction. Chemistry recall, arithmetic expressions, explicit logic, and Python output prediction differ in length, vocabulary, answer type, and semantic diversity. Because joint length matching failed and family counts are small, this experiment does not identify which property produces the contrast.

Spectral and local estimates also need separate interpretation. A cloud can distribute variance unevenly and still have nontrivial local neighbor ratios. Conversely, curvature and sparse sampling can separate global covariance geometry from local dimension. With only 12 points, disagreement is not evidence for a particular manifold mechanism. The two-dimensional projection in the appendix is a visualization fitted on exploratory data, not a dimensionality test.

Our experiment measures input-dependent representations before any generated answer. It supplies no direct evidence that a larger effective dimension means more accurate reasoning, increased computational effort, or a better language model. Such hypotheses require additional behavioral measurements and interventions.

10 Limitations and future directions

The principal limitations are one pretrained base model, one modest model size, 12 held-out prompts and two held-out families per category, and no supported joint length control. The finite covariance rank strongly constrains the spectral estimates, while sparse neighborhoods limit TWO-NN reliability. Named families are operational clusters, not guaranteed independent samples from a well-defined population. The restricted chemistry and Python-expression tasks do not cover factual knowledge or software development broadly.

Last-token pooling is another scope choice. The shared final suffix standardizes the last input token but may influence contextual geometry. Different pooling positions, paraphrases, matched answer formats, and alternative sources could change the result. Hardware and precision are documented, but warning-only deterministic settings do not promise bitwise reproducibility across environments. The archive lacks activation arrays; exact reanalysis requiring vectors must rerun extraction.

Future work should increase independent family diversity, improve tokenizer-level length overlap before extraction, review task ambiguity, and compare additional models and pooling choices. These are recommendations rather than experiments completed in this paper. Resource limitations explain the present scale but do not justify stronger inference than the data support.

11 Conclusion

A 200-prompt exploratory study of Qwen2.5-1.5B reveals depth-dependent differences between task-category representation clouds. In the 12-prompt-per-category held-out analysis, the reasoning–recall PR contrast changes from positive at blocks 1–20 to negative at blocks 21–28. The result is descriptive and remains limited by family counts, sample rank, and insufficient length overlap. The paper contributes a reproducible comparison and its analysis artifacts, without claiming a general task ordering or a causal account of reasoning.

Data and code availability

Our project repository is `https://github.com/mitravarun123/Interensic-Dimenalisty-`. Released numerical CSVs are available at `https://github.com/mitravarun123/Interensic-Dimenalisty-/tree/main/recovered_results/qwen_pilot_200%202/results`. We license the author-owned manuscript sources and figures under the MIT license in `LICENSE`; third-party components, datasets, and model weights retain their respective licenses. Activation arrays are not released; rerun extraction to regenerate.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of ACL-IJCNLP*, pages 7319–7328, 2021. doi: 10.18653/v1/2021.acl-long.568. URL `https://aclanthology.org/2021.acl-long.568/`.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL `https://papers.neurips.cc/paper_files/paper/2019/hash/cfcce0621b49c983991ead4c3d4d3b6b-Abstract.html`.

Emily Cheng, Diego Doimo, Corentin Kervadec, Iuri Macocco, Jade Yu, Alessandro Laio, and Marco Baroni. Emergence of a high-dimensional abstraction phase in language transformers. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2405.15471.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, pages 55–65, 2019. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006/.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7:12140, 2017. doi: 10.1038/s41598-017-11873-y. URL https://www.nature.com/articles/s41598-017-11873-y.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202.

Luca Mazzucato, Alfredo Fontanini, and Giancarlo La Camera. Stimuli reduce the dimensionality of cortical activity. *Frontiers in Systems Neuroscience*, 10:11, 2016. doi: 10.3389/fnsys.2016.00011.

Qwen Team. Qwen2.5 Technical Report, 2024. URL https://arxiv.org/abs/2412.15115.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference*, pages 606–610, 2007. URL https://doi.org/10.5281/zenodo.40328.

A Prompt inventory and family construction

The accompanying `data/prompts_200.jsonl` supplies prompt identifiers, categories, family labels, partitions, token lengths, fixtures, and verification labels for all 200 prompts. Table 4 shows one actual example per family. Chemical-element facts use fixed chemistry fixtures; Arithmetic, logic, and Python fixtures are recorded as computed values. Their presence does not certify model knowledge or model answer accuracy. Family-level partitioning prevents the same named family from appearing in both partitions, but shared entities, operators, and vocabulary can still cross partitions.

Table 4: One actual prompt per family. E: exploratory; H: held-out. All 200 prompts and answer fixtures are recorded in the source archive.

Category/family	Split	Prompt	Answer
Factual recall 0	E	What is the chemical symbol for hydrogen? Reply with only the answer.	H
Factual recall 1	E	Which element has atomic number 2? Reply with only the answer.	helium
Factual recall 2	E	What is the atomic number of lithium? Reply with only the answer.	3
Factual recall 3	E	Name the element represented by Be. Reply with only the answer.	beryllium
Factual recall 4	E	Recall the symbol used for the element boron. Reply with only the answer.	B
Factual recall 5	E	Identify the element at position 6 in the periodic table. Reply with only the answer.	carbon
Factual recall 6	H	State how many protons a neutral nitrogen atom has. Reply with only the answer.	7
Factual recall 7	H	Which element is denoted by the chemical symbol O? Reply with only the answer.	oxygen
Arithmetic 0	E	What is 99 + 16? Reply with only the answer.	115
Arithmetic 1	E	Calculate 23 multiplied by 89. Reply with only the answer.	2047
Arithmetic 2	E	What is 2618 divided by 34? Reply with only the answer.	77
Arithmetic 3	E	If x = 74, what is 75x + 5? Reply with only the answer.	5555
Arithmetic 4	E	Subtract 5 from 30. Reply with only the answer.	25
Arithmetic 5	E	Compute (75 + 19) times 2. Reply with only the answer.	188
Arithmetic 6	H	What is the remainder when 610 is divided by 29? Reply with only the answer.	1
Arithmetic 7	H	Square 50 and subtract 51. Reply with only the answer.	2449
Reasoning 0	E	Jules is older than Charlie. Charlie is older than Hugo. Who is oldest? Reply with only the answer.	Jules
Reasoning 1	E	All Grace objects are Kiran objects. All Kiran objects are Iris objects. Are all Grace objects Iris objects? Reply with only the answer.	yes
Reasoning 2	E	An alarm activates exactly when both sensors trigger. Sensor Diana is false and sensor Iris is true. Does it activate? Reply with only the answer.	no
Reasoning 3	E	Hugo is left of Bob, and Bob is left of Kiran. Who is furthest right? Reply with only the answer.	Kiran
Reasoning 4	E	The lamp is on exactly when at least one switch is on. Switch Kiran is false and switch Fatima is true. Is the lamp on? Reply with only the answer.	yes
Reasoning 5	E	Hugo has more coins than Priya. Fatima has fewer coins than Priya. Who has the fewest coins? Reply with only the answer.	Fatima
Reasoning 6	H	A gate opens exactly when one of two inputs is true, but not both. Input Lina is true and input Hugo is true. Does it open? Reply with only the answer.	no
Reasoning 7	H	Maya finished before Noah; Noah finished before Fatima. Who finished second? Reply with only the answer.	Noah
Coding 0	E	In Python, what is sum([39, 6, 3, 20, 14])? Reply with only the answer.	82
Coding 1	E	What does Python max([3, 2, -7, -2, -19]) return? Reply with only the answer.	3

Category/family	Split	Prompt	Answer
Coding 2	E	What is the output of 'igjfki'[::-1] in Python? Reply with only the answer.	ikfjgi
Coding 3	E	What does sorted([34, 12, 40, -15, -13]) return in Python? Reply with only the answer.	[-15, -13, 12, 34, 40]
Coding 4	E	What is len('hdabda') in Python? Reply with only the answer.	6
Coding 5	E	What is the output of [x * 2 for x in [12, 2, 48, -4, 8]]? Reply with only the answer.	[24, 4, 96, -8, 16]
Coding 6	H	What does min([-18, -17, 6, 46, -17]) return in Python? Reply with only the answer.	-18
Coding 7	H	What does list(reversed([36, 6, -17, 43, -4])) return in Python? Reply with only the answer.	[-4, 43, -17, 6, 36]

B Length distribution

Category	Min	Q1	Median	Q3	Max
Factual recall	13	14.25	16	17	20
Arithmetic	15	16	17	19	23
Reasoning	21	24.25	29	31	38
Coding	17	28.5	31	33	40

Table 5: Exact tokenizer-length quantiles across all 50 prompts per category. These are full-dataset summaries, not held-out-only quantiles.

The analysis attempts to equalize counts within common four-token bins. Its recorded outcome is `no_common_length_bins`, with no shared bins and zero matched prompts. This is a failed feasibility condition, not a negative result from a length-controlled experiment.

C Numerical implementation and validity checks

Economy SVD operates on the centered float64 matrix. Numerical rank uses singular-value tolerance $\tau = \max(N, d)\epsilon_{64}s_1$ by default. This tolerance diagnoses numerical rank; it does not truncate the positive spectrum before PR or entropy computation. Very small positive residual eigenvalues may therefore remain in spectral plots and entropy calculations. PR rescales eigenvalues by their maximum for numerical stability. Entropy excludes exactly zero probabilities. Substantial negative covariance eigenvalues are rejected; tiny negative values in a recorded spectrum are clipped within a floating-point tolerance.

Tests in the software cover mathematical identities, degenerate inputs, known synthetic structures, extraction bookkeeping, and analysis workflow behavior. Such tests validate software behavior rather than the scientific sampling assumptions. No synthetic result is substituted for a Qwen observation in this paper. TWO-NN's ten-point minimum means eight-point subsamples cannot support its fit even when spectral summaries can be computed.

D Complete spectral results

Table 6: Complete held-out raw spectral estimates. Each category uses 12 prompts; ER denotes covariance effective rank.

Block	Factual recall PR	ER	Arithmetic PR	ER	Reasoning PR	ER	Coding PR	ER
1	1.066	1.225	1.038	1.130	1.160	1.478	2.637	3.764

Block	Factual recall PR	ER	Arithmetic PR	ER	Reasoning PR	ER	Coding PR	ER
2	1.095	1.315	1.040	1.144	1.255	1.715	2.475	3.626
3	1.187	1.557	1.060	1.204	1.462	2.179	2.268	3.391
4	1.234	1.676	1.048	1.170	1.547	2.355	2.172	3.237
5	1.279	1.788	1.056	1.196	1.601	2.476	2.105	3.170
6	1.240	1.707	1.075	1.251	1.624	2.601	1.969	2.992
7	1.267	1.776	1.091	1.297	1.614	2.609	1.935	2.923
8	1.247	1.732	1.094	1.306	1.509	2.388	1.842	2.749
9	1.193	1.577	1.105	1.340	1.433	2.199	1.692	2.514
10	1.179	1.540	1.110	1.352	1.389	2.086	1.532	2.222
11	1.174	1.534	1.123	1.389	1.373	2.056	1.431	2.045
12	1.143	1.451	1.134	1.419	1.384	2.085	1.350	1.879
13	1.165	1.508	1.143	1.446	1.359	2.023	1.396	1.967
14	1.169	1.514	1.145	1.453	1.324	1.936	1.375	1.938
15	1.165	1.498	1.171	1.521	1.301	1.879	1.302	1.781
16	1.147	1.449	1.171	1.527	1.267	1.789	1.272	1.721
17	1.124	1.381	1.217	1.648	1.260	1.752	1.259	1.696
18	1.132	1.402	1.231	1.682	1.212	1.631	1.278	1.736
19	1.137	1.417	1.255	1.745	1.178	1.543	1.295	1.786
20	1.119	1.367	1.306	1.878	1.183	1.549	1.300	1.797
21	1.289	1.783	1.345	1.991	1.151	1.465	1.327	1.875
22	1.420	2.041	1.332	1.963	1.184	1.562	1.323	1.874
23	2.000	3.102	1.334	1.967	1.245	1.720	1.324	1.867
24	2.869	4.152	1.325	1.947	1.397	2.064	1.323	1.857
25	2.779	4.086	1.299	1.880	1.640	2.583	1.326	1.863
26	2.669	3.973	1.299	1.879	1.623	2.560	1.337	1.880
27	2.784	4.255	1.294	1.868	1.913	3.155	1.306	1.808
28	2.235	3.689	1.369	2.036	2.126	3.624	1.354	1.923

All PCA thresholds, TWO-NN values and flags, alternate preprocessing modes, and all six category-pair contrasts are recorded as machine-readable CSVs in `data/`. The printed table is restricted to PR and entropy effective rank to avoid duplicating thousands of tabular entries. Index 0 is omitted here because its primary clouds have zero variance. Index 28 includes final RMSNorm.

E Preprocessing and uncertainty checks

Block	Raw	Unit norm	Standardized
6	+0.384	+0.392	+0.853
20	+0.064	+0.057	+0.229
21	-0.138	-0.153	-0.360
24	-1.472	-1.363	-1.773
28	-0.110	-0.244	-0.300

Table 7: Selected reasoning-minus-recall PR contrasts under three preprocessing choices, held-out sample-matched subsets, $N = 12$ per category.

F Additional figures

The five figures below retain original outputs not needed as separate main-text figures. Together with the two figures in the main text, they preserve all seven figure families. Two additional paper-specific main figures were reconstructed from the original comparisons and prompt records; they introduce no new model run.

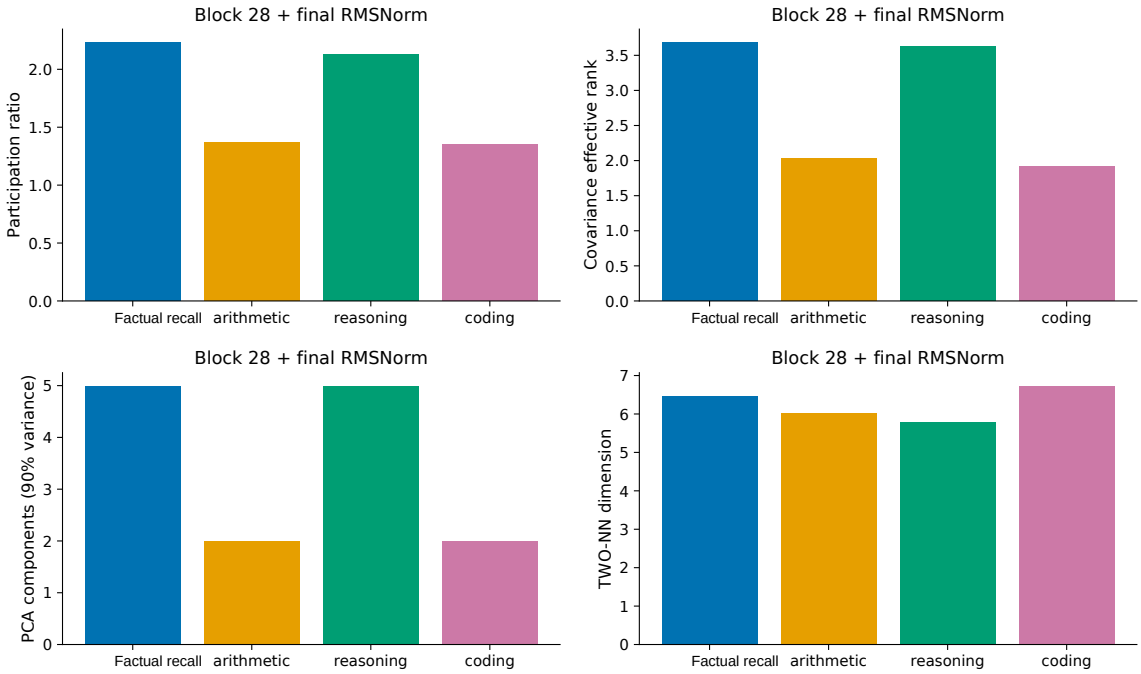


Figure 5: Original final-output category comparison. The values duplicate Table 3 graphically and are retained here for reference. Raw held-out representations, $N = 12$ per category, output 28 including final normalization.

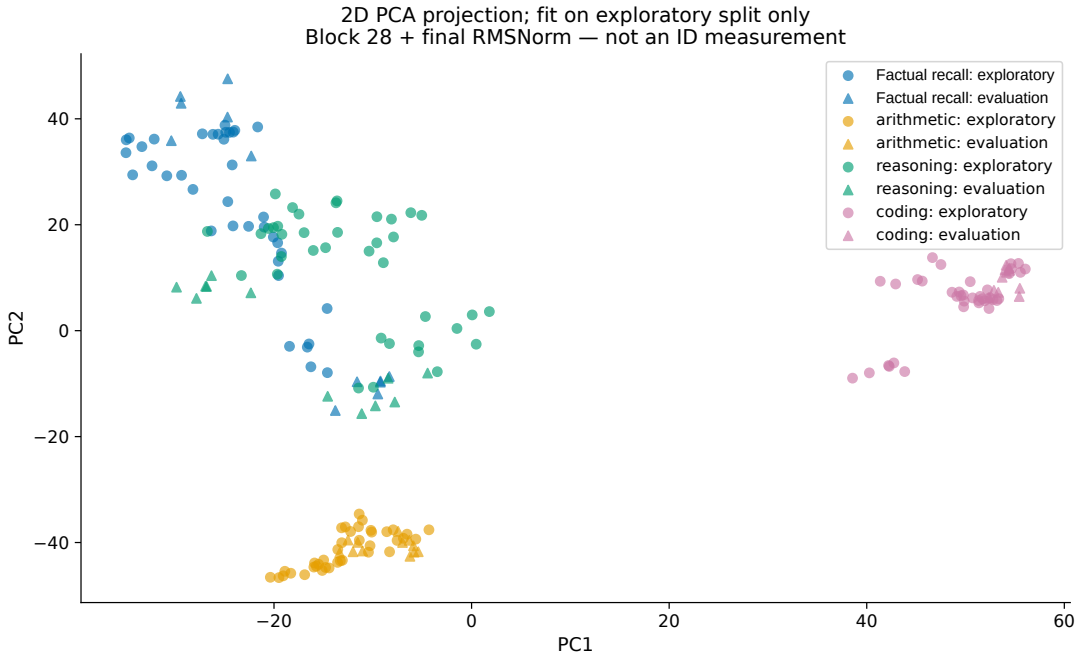


Figure 6: Original two-dimensional PCA projection at output 28. The PCA basis is fitted on the exploratory partition only; markers distinguish exploratory and held-out prompts. Category separation in this display is not an intrinsic-dimension measurement or a test of task effects.

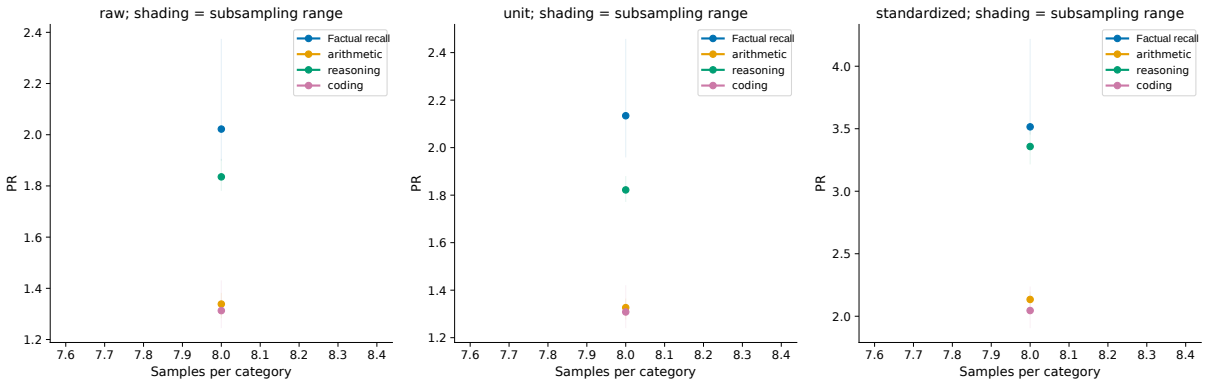


Figure 7: Original final-output PR subsampling panel under three preprocessing modes. Shading is the range over repeated subsets, not a confidence interval. Only eight-point subsets are feasible from the 12-point held-out category clouds; requested larger subsets are unavailable.

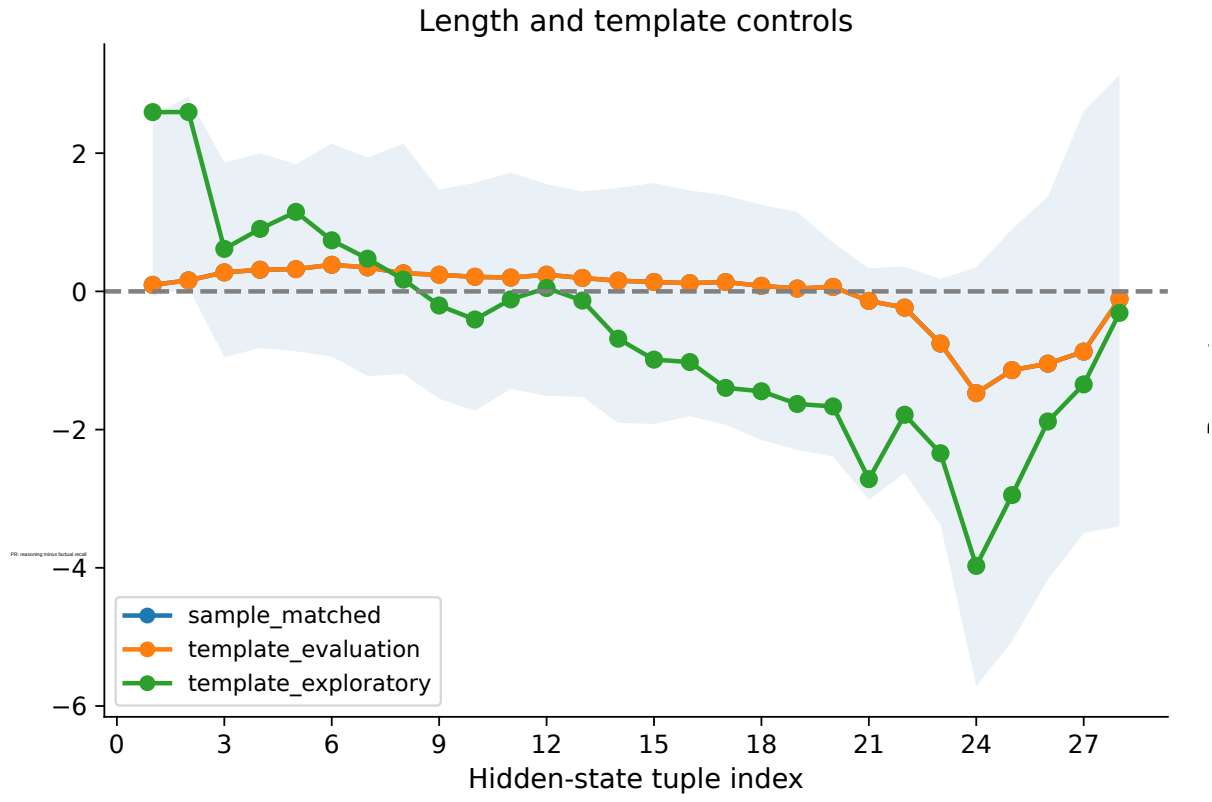


Figure 8: Original template and length checks. The left panel compares different subsets; any available exploratory intervals refer to that partition, not the primary held-out contrast. The right panel uses all prompts and reports unsuccessful joint length matching. Sample counts and families differ between exploratory and evaluation curves.

Quantity	Availability and interpretation
Primary family-bootstrap interval	Withheld: two held-out families per category, below the configured minimum of four.
Primary permutation p-value	Not computed: family exchangeability not assumed.
Adjusted primary p-value	Not computed because no primary p-value is available.
Embedding-layer spectrum	Zero variance; normalized spectral dimension undefined.
TWO-NN bootstrap interval	Disabled because replacement resampling produces duplicate points.
Joint matched-length contrast	Not computed: no common bins across the four categories.
Subsampling spread	Descriptive range across repeats, not an inferential interval.

Table 8: Reasons for unavailable quantities. These cases must not be interpreted as equivalent forms of a null result.

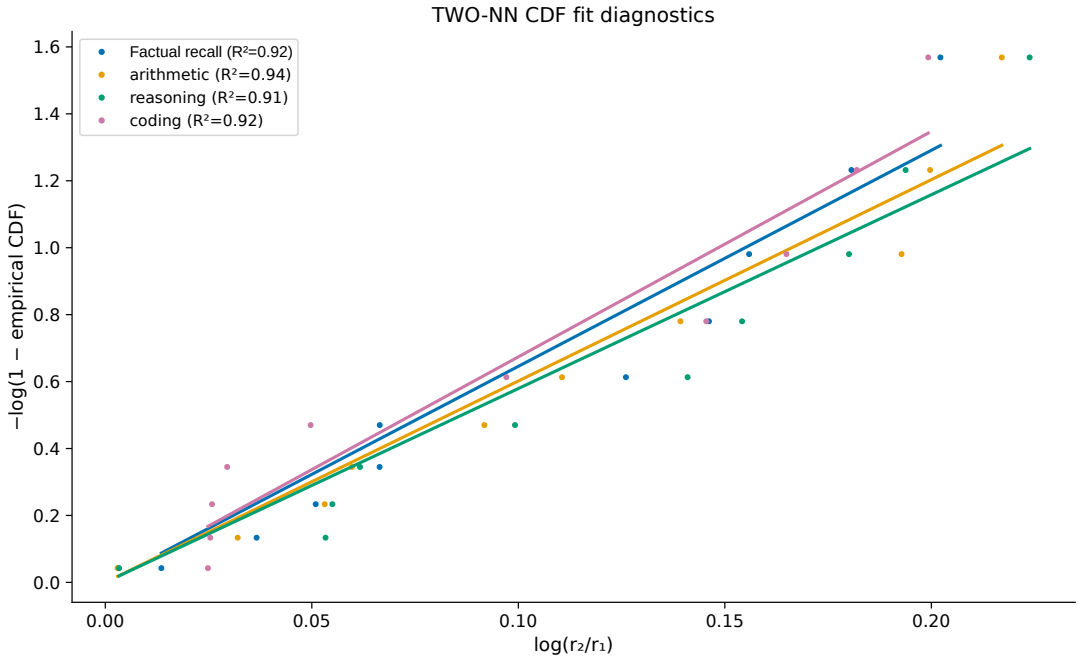


Figure 9: Original TWO-NN empirical-CDF fits at output 28. Each category contributes only ten retained fit points from 12 prompts. Reported R^2 values assess this fit, not the validity of manifold assumptions or estimator accuracy.

G Environment and reproducibility

Item	Recorded value
Model	Qwen/Qwen2.5-1.5B (base)
GPU	Tesla T4
Extraction dtype / batch size	float32 / 1
Analysis arithmetic	float64
Seed	42
Python	3.13.15
PyTorch	2.11.0+cu130
Transformers	4.46.3
Tokenizers	0.20.3
NumPy / SciPy	2.1.3 / 1.16.3
Matplotlib	3.10.0
Hugging Face Hub	0.36.2
Tokenizer	Qwen2TokenizerFast
Padding side / pad ID	right / 151643
Attention	eager
Deterministic algorithms	warn-only

Table 9: Values recorded in the extraction metadata; these describe the model run, not the machine used to typeset this paper.

The initial requested revision was unspecified; the resolved revision is recorded in Section 4 and should be pinned for reruns. The analysis fingerprint is `d0e85adceeec469cde778b734e863203c2c35b8a903e24bf92cf67cab5aac003`. Reproduction requires keeping prompt order, split labels, preprocessing, and estimator settings fixed. Some environment fields, such as a standalone GPU driver version, are not recorded and are not recorded.

Activation arrays are not released; rerun extraction to regenerate.