
Entanglement Is All You Don’t Need

Wei Sheng

Mandelbrot Technology Co., LTD,
Shenzhen, China
shengwei@mandelbrot.cn

Chengzhu Xiao

Lunhao Ao

JunYan Long

Ye Yu

Abstract

This paper primarily focuses on mechanistic investigation, conducting controlled experimentation at a 1B parameter scale to effectively isolate scaling effects, rather than large language model optimization aimed solely at achieving higher benchmark scores. The intrinsic dimensionality collapse observed in DeepSeek-V2-Lite-Chat, Gemma, and Qwen indicates premature feature fusion, leading to feature entanglement, which limits the model’s representational capacity. We believe that feature entanglement due to ID collapse may be an important factor contributing to the Scaling Law ceiling. To address this problem, we propose a principled architectural design called EPMORE (Explainable Process Mixture-of-Experts) based on the core intuition that the reasoning process is a conditionally dependent feature (from abstract to concrete) accumulation process. EPMORE consists of four major components. SPE (Simple Position Encoding) achieves disentanglement between position and content. EPL (Expansive Processing Logic) increases feature capacity by dynamically expanding from extremely low dimensions, making feature disentanglement possible. PIL (Parameter Independence Loss) in the architecture decouples the factor entanglement in weight matrices, thereby achieving feature disentanglement. The MOR (Middle Output Reuse) mechanism allows tokens whose intrinsic dimensionality increases during inference to align with tokenizers at different levels of abstraction; during training, intermediate ground truth labels (not manually annotated) aligned in a hierarchical manner based on the token’s level of abstraction (i.e., intrinsic dimensionality) provide intermediate losses that shorten backpropagation paths, accelerating training and reducing computational cost. Ultimately, it enables feature disentanglement between layers. Experiments demonstrate that the EPMORE architecture alleviates feature entanglement, making it possible to ultimately achieve interpretability and break through the bottlenecks of the scaling law. The example of “the capital of France” in the supplementary materials clearly shows how the model derives the result step by step according to the semantic hierarchy.

1 Introduction

Imagine an AI whose thinking resembles a chaotic mass of muddled paste — vast yet hopelessly entangled. This is the reality facing many of today’s large language models. Despite exponential growth in model parameter counts, the problem of feature entanglement caused by intrinsic dimensionality collapse — a phenomenon clearly observed in three benchmark models of the same parameter scale and trained on the same dataset (see **Figures 2, 7, and 8**) — leaves models struggling to handle complex reasoning tasks. This entangled representational state may contribute to the diminishing marginal returns predicted by Scaling Laws. Most existing solutions attempt to brute-force their

way through by stacking ever more data or parameters, yet few works fundamentally reconstruct the logical structure of the reasoning process itself.

We firmly believe that the next generation of AI should think with clarity and precision. To quantify this gap, we use intrinsic dimensionality as the metric: the closer the intrinsic dimensionality is to the physical dimensionality, the closer the representation is to the theoretical ideal.

Guided by this principle and adhering to minimal invasiveness without modifying the fundamental Transformer operators, we propose EPMORE (Explainable Process Mixture of Experts). Unlike conventional models, EPMORE does not pursue blind scaling; instead, it focuses on feature disentanglement during the reasoning process. To validate this goal, we design two sets of experiments.

Compared with multiple baseline models during pretraining, under the same parameter scale and on the same dataset, we conduct pretraining from scratch and examine the changes in intrinsic dimensionality at the input and output of each block, as well as the quantitative metric of parameter independence. As shown in Figures 1 and 2 (with results aggregated over ten independent runs to ensure statistical robustness): the intrinsic dimensionality of EPMORE increases with module index, and in most modules the input intrinsic dimensionality is lower than the output intrinsic dimensionality, indicating that EPMORE achieves intrinsic-dimensionality expansion during inference, that is, feature disentanglement occurs. By contrast, under the same strict evaluation protocol, several benchmark models consistently exhibit non-independent feature dimensions, leading to reproducible geometric collapse in intrinsic dimensionality, a clear manifestation of feature entanglement. This behavior is unaffected by random initialization or data sampling. Ultimately, the number of independent features is constrained, and the model’s representational capacity is fundamentally bottlenecked by the fixed-width architecture.

By designing a reasoning case centered on the question “What is the capital of France?”, we examine the consistency of EPMORE’s hierarchical reasoning and highlight the potential of process-level interpretability.

In the linked video, an example illustrating the feature disentanglement phenomenon during EPMORE inference is presented: the example of the capital of France (See Figure 3 and the video of the capital of France is available on the website at <https://doi.org/10.5281/zenodo.19342615>).

We have open-sourced the experimental code to support reproducibility research and to provide a foundation for structured, interpretable language model research.

The training code for these comparative experiments on the dataset on Hugging Face can be found in Supplementary Material A.

The inference code for the example of the capital of France is in Supplementary Material B.

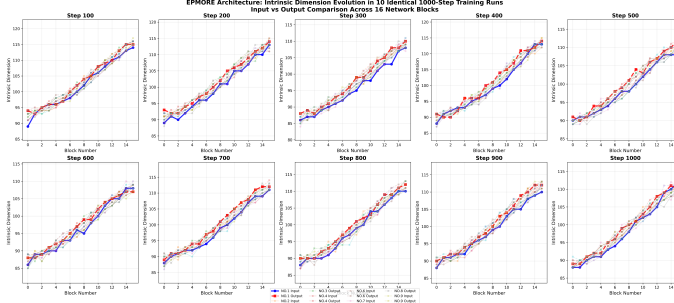


Figure 1: EPMORE Intrinsic Dimensionality Progression

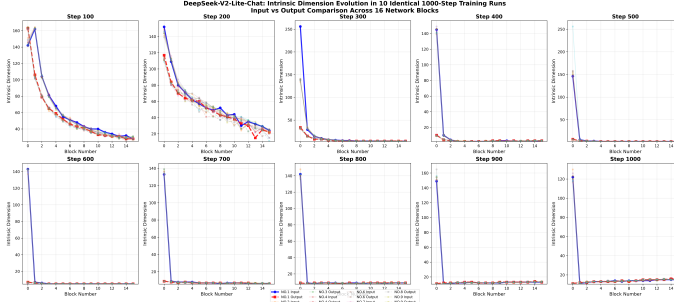


Figure 2: DeepSeek-V2-Lite-Chat Intrinsic Dimensionality Collapse

Please refer to Figure 9, 10, 11, 12 for the t-Test reports of the intrinsic dimension expansion and collapse experiments.

2 Methodology

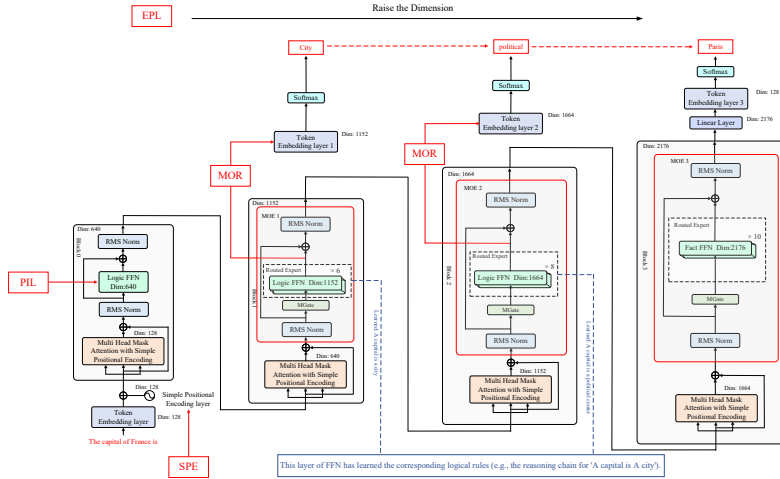


Figure 3: The Architecture of The Example of The Capital of France

2.1 EPMORE Architecture Overview

Mathematical Intuition and Operator Definition – To help readers understand the author’s design philosophy, rather than to claim that the empirical model fully achieves them.

First, we provide an intuitive mathematical perspective to understand the core principles of EPMORE. EPMORE’s essence—dimension elevation, hierarchical self-similarity, and feature disentanglement—can be understood analogously to the imaginary unit i in complex numbers. Let $\mathbf{x} \in \mathbb{R}^n$

be a low-dimensional feature vector. We define a dimension-elevating operator $\mathcal{E}$ and a feature disentanglement operator $\mathcal{D}$ as follows:

$$\mathbf{z}^{(L)} = (\mathcal{D} \circ \mathcal{E})^L(\mathbf{x}), \quad \mathcal{E}(\mathbf{x}) = \mathbf{x} + i\mathbf{x}_\perp \quad (1)$$

Here, $\mathcal{E}$ acts like multiplication by i : it elevates the feature dimension while preserving the original structure. $\mathcal{D}$ disentangles features, analogous to separating amplitude and phase. L denotes the number of hierarchical blocks, implementing iterative self-similar expansion and disentanglement.

Intuition: EPMORE constructs a hierarchical, interpretable feature space where low-dimensional entangled features are iteratively lifted and disentangled. Each EPL block functions as a feature-space analog of a self-similar, structure-preserving dimension-elevating operator, just like i in complex numbers.

Transition to formal proofs: The subsequent sections rigorously establish injectivity, structure preservation, and other properties of $\mathcal{E}$ and $\mathcal{D}$. This analogy serves as an intuitive guide, without being required for the formal proofs.

EPMORE organizes the model into two functional domains:

- **Hierarchical Logical Space** (Expand hierarchically from higher to lower levels of abstraction, with dimensions increasing from extremely low to high.)
- **Factual Space** (high-dimension, MoE)

These regimes serve as an architectural prior for structuring feature transformations. Within each domain, ReLU-induced sparsity promotes parameter decoupling and modular computation. Combined with our self-attention mechanism and simple positional encoding, these components together constitute the architecture of EPMORE.

2.2 Simple Position Encoding (SPE)

A dedicated dimension is appended to the hidden state to encode position independently from content information. SPE achieves decoupling of positional and content representations, simplifies the scaling trajectory, and reduces computational overhead compared to RoPE, while preserving the positional awareness required for downstream processing. Positional encoding is also designed for information disentanglement.

2.3 Expansive Processing Logic (EPL) – See Appendix A

The precise definition of EPL is: a hierarchical dimensionality-lifting operator followed by a ReLU activation. EPMORE adopts monotonically increasing dimensionality across layers while incorporating Expansive Processing Logic. The dimensionality expansion phase, by providing a high-capacity representational space, facilitates the decoupling of heterogeneous logical features and reduces representational entanglement, thereby enhancing model performance.

2.4 Parameter Independence Loss (PIL) – See Appendix B

Let $A_k(x)$ denote the active neurons for input x at layer k . The total number of neurons in layer k is d_{ff} . Then, the Parameter Independence Loss metric:

$$I_k(x, y) = \frac{|A_k(x)|}{d_{\text{ff}}} + \frac{|A_k(y)|}{d_{\text{ff}}} - \frac{|A_k(x) \cap A_k(y)|}{d_{\text{ff}}} \quad (2)$$

2.5 Middle Output Reuse (MOR) – See Appendix C

Middle Output Reuse Mechanism (MOR) is the core mechanism of EPMORE for implementing empirical analysis and inductive bias during the training phase, designed to overcome the limitations of post-hoc probing. MOR is not an interpretability method, nor does it prescribe target representations; rather, it is a diagnostic and alignment interface that exposes intermediate layer outputs to the same token-level prediction interface as the final layer. MOR serves two mutually exclusive roles (user-configurable):

2.5.1 Diagnostic tool for the inference phase

During inference, MOR reuses intermediate hidden states $(h^{(2)}, h^{(3)}, h^{(4)})$ to generate token-level predictions $(y^{(2)}, y^{(3)}, y^{(4)})$ without modifying model parameters.

2.5.2 Inductive bias during the training phase

During training, MOR can optionally provide a mild auxiliary gradient signal by incorporating the loss from intermediate predictions:

$$L_{\text{total}} = L_{\text{final}} + \alpha L_{\text{intermediate}} \quad (3)$$

Where $\alpha \ll 1$ (a small weight, ensuring the final task remains the primary optimization objective). These auxiliary signals serve as a form of inductive bias, achieving the following effects:

- Preventing premature information collapse (the model cannot compress all task-relevant information into later layers);
- Encouraging hierarchically specialized representations (early/intermediate layers must encode reusable features applicable to the task in order to minimize intermediate loss);
- Stabilizing the training dynamics of sparse MoE layers, reducing gradient starvation in low-activation experts.

The core idea of MOR during the training phase is: using the level of abstraction of tokens (quantified in this paper by intrinsic dimensionality) as the basis, constructing hierarchically aligned intermediate ground-truth labels for the model’s intermediate layers, and computing intermediate-layer losses from these labels to provide stable, semantically meaningful gradient signals to intermediate layers — thereby achieving layer-wise supervision and feature disentanglement.

2.6 Fact-level MoE (Mixture of Experts) layer

Location: Stage 4 of the architecture, at the dimensional peak (2176 dimensions);

Function: Grounds the preceding logical reasoning results into concrete facts;

Mechanism: Sparse MoE, where each expert is responsible for a specific factual sub-domain;

Advantage: High-dimensional space reduces representational entanglement between facts; expert isolation prevents knowledge confusion;

Interpretability: Via MOR, the vocabulary-level prediction output of this layer can be directly observed.

In short, the "fact-level MoE layer" is the core hub for knowledge storage and fact retrieval in EPMORE. It sits at the endpoint of the reasoning chain, responsible for converting the abstract "semantic + logical" combination of objects that have been reasoned through to the final stage into concrete factual answers.

The "fact-level MoE layer" is the fourth stage in the architecture diagram — the Mixture-of-Experts layer at the dimensional peak (2176 dimensions). The paper uses "the capital of France" as an illustrative example, presenting a schematic table (Table 2). It can be observed that the preceding layers handle abstract relations and attributes (city, French, political center...), while the MoE layer handles the combination of these attributes to lock onto a concrete fact (Paris). This is the origin of the name "fact-level": this layer is responsible for grounding the results of logical reasoning into specific, retrievable factual answers.

Table 1: Structural Disentanglement: Levels and Axioms

Level	Disentanglement Object	Core Problem	EPMORE Mechanism
Information Type	Position $\perp$ Semantics	Positional encoding pollutes the semantic space	SPE
Representation Space	Abstract $\perp$ Concrete	General and specific features compete for dimensions	EPL
Computation Path	Feature A $\perp$ Feature B	Overlapping neuron activations cause interference	PIL
Supervision Target	Level $k \perp$ Level k'	Intermediate supervision requires manual design	MOR

3 Experiments

Disentanglement does not have a universally agreed-upon definition, particularly in the context of language models where ground-truth generative factors are not observable. In this work, we adopt an **operational definition of disentanglement** based on measurable properties of representations. Specifically, we characterize disentanglement through:

- reduced activation overlap between inputs (parameter independence),
- increased intrinsic dimensionality (avoiding representational collapse),
- and structured, hierarchical intermediate representations.

We emphasize that this definition captures a **practical and observable notion of disentanglement**, rather than an idealized factorized latent representation.

Furthermore, our experiments are controlled experiments, and their purpose is not to inflate scores on benchmarks.

3.1 Intrinsic dimensionality experiments

The intrinsic dimensionality experiment was conducted using the GSM8K dataset at the 1B parameter scale, with experiments performed separately on DeepSeek-V2-Lite-Chat, Gemma4-E4B, Qwen2.5 and EPMORE. See Table 4 for the data of intrinsic dimensionality experiments.

3.1.1 Experimental objective

Large Language Model Inference Mechanism Research: to demonstrate that DeepSeek-V2-Lite-Chat exhibits feature entanglement during inference, whereas EPMORE exhibits feature independence (i.e., non-entanglement) during inference.

3.1.2 Algorithm

Intrinsic Dimensionality:

For each block, by taking the tensors corresponding to inputs_ID and outputs_ID, applying PCA, and sorting the singular values in descending order, the intrinsic dimensionality is defined as the maximum value of k such that the sum of squares of the top- k singular values divided by the sum of squares of all singular values is less than the threshold of 0.95.

3.1.3 Experimental setup

Evaluation Metrics:

- **Sparsity** (Nonzero activation rate)
- **Parameter independence** (the intersection of the two sets of neurons activated by two pieces of text in each FFN layer)
- **Efficiency** (FLOPs, floating-point operations count, and throughput)
- **Performance** (loss / perplexity)

3.1.4 Experimental results

As shown in Figure 1, the intrinsic dimensionality of EPMORE did not collapse during the experiment; as shown in Figure 2, the intrinsic dimensionality of DeepSeek-V2-Lite-Chat did.

- **Sparsity:** EPMORE exhibits significantly higher activation sparsity.
- **Parameter Independence:** Activation overlap across tokens is reduced, indicating more decoupled computational pathways.
- **Efficiency:** Lower FLOPs and higher throughput.
- **Performance:** On par with or improved compared to the baseline.

Conclusion: In the intrinsic dimensionality experiments, it reduced activation overlap and increased intrinsic dimensionality.

3.2 Reasoning experiment on the capital of France

The architecture diagram of the reasoning process of EPMORE is provided in Figure 3.

The results for the experiments on Ablation-MOR and Ablation-EPL can be found in Supplementary Material B.

A video of the reasoning process of the capital of France is available on the website at <https://doi.org/10.5281/zenodo.19342615>.

See Table 5 for the experimental data of the capital of France.

3.2.1 Experimental objective

- **Objective 1: Validating the Diagnostic Function of MOR**

The core claim of MOR is that "the outputs of intermediate layers can be directly reused for vocabulary-level prediction." This example is intended to demonstrate whether, during inference, intermediate layers at different depths truly correspond to semantic states of varying granularity, and whether each layer can produce meaningful vocabulary predictions.

- **Objective 2: Visualizing the Process of "Alleviating Feature Entanglement"**

The core claim of EPMORE is that representations across layers gradually specialize and decouple as depth increases. This example attempts to translate this abstract technical claim into a human-interpretable form using an intuitive natural language case — specifically, illustrating the stepwise convergence from "a city" → "a city in France" → "the political center of France" → "Paris."

3.2.2 Experimental setup

Input: "The capital of France is..."

Observation mechanism: MOR outputs layer by layer.

The specific process can be referred to in Table 2 .

Table 2: The reasoning process and hierarchical prediction using the example of the capital of France

Level	Predicted Output	Tensor Semantics
Logic Layer 1	"city"	Base concept vector V_{city}
Logic Layer 2	"political center"	Superimposed "capital" attribute
Factual MoE Layer	"Paris"	Composed into a concrete factual entity

Ablation Comparison from Multiple Perspectives

This example is also embedded within an ablation study framework, corresponding to the following three comparison conditions:

- **EPMORE (Full version):** Includes EPL (Expansion Processing Logic) + MOR.
- **EPMORE (Remove MOR):** Architecture expansion only, without intermediate supervision signals.
- **EPMORE (Remove EPL):** No dimensional expansion, degenerates into a fixed-dimension structure.

By comparing the intermediate prediction quality and final loss across these three conditions, the contribution of each component is evaluated.

3.2.3 Experimental Summary

Direct Observation Conclusions:

The direct observations drawn from this example in the paper are:

The intermediate predictions across layers exhibit a progressive convergence from coarse-grained to fine-grained, consistent with the hierarchical structure described in the EPMORE framework:

- **Hierarchical Logical layers** predict restrictive attributes ("city", "political center").
- **Factual MoE layers** predict specific factual answers ("Paris").

Conclusions of the reasoning experiment on the capital of France: In the EPMORE architecture, layer-wise consistency was enhanced. **Conclusions of the Ablation Study:**

Table 3: Ablation Study Results of EPMORE Components

Condition	Observations
Remove EPL (no dimensional expansion)	loss=0.2679, ppl=1.31, worst intermediate representation organization, worst final convergence
Remove MOR (no intermediate supervision)	loss=0.0231, ppl=1.02, faster initial training decline, longer total training time, but weaker representation space organization than when EPL is removed
Full EPMORE	loss=0.0203, ppl=1.02, optimal convergence, most ordered representations

Removing EPL leads to the greatest performance degradation; removing MOR causes performance to degrade more slowly than in the full model but faster than when EPL is removed, while also resulting in weaker representation organization; the full model achieves the best overall performance.

the total conclusion of the two experiments: ¹ Our empirical results demonstrate a strong cross-layer correlation between reduced activation overlap, elevated intrinsic dimensionality, and enhanced hierarchical consistency. We propose that these metrics are diverse manifestations of a unified underlying phenomenon: the emergence of structured feature disentanglement. Furthermore, by quantifying the gap between intrinsic dimensionality and physical dimensionality, EPMORE demonstrates that its structural intervention effectively alleviates feature entanglement — the closer the intrinsic dimensionality approaches the physical dimensionality, the closer the representations approach the IHSR theoretical upper bound.

4 Related Work

4.1 Mixture-of-Experts (MoE)

MoE models improve efficiency through sparse routing, but typically operate within this same fixed-dimensional space. Sparse architectures, and mixture-of-experts (MoE) models in particular ([14]; [4]), improve efficiency by activating only a subset of parameters. However, most conventional MoE approaches focus on computational savings at a uniform scale.

4.2 Interpretability research

Mechanistic interpretability methods analyze already-trained models — that is, they attempt to decode the "black box" nature of large language models through post-hoc analysis of neurons and circuits — but do not enforce interpretability at the architectural level. In contrast, EPMORE directly embeds process-level interpretability into the architecture through Middle Output Reuse (MOR). EPMORE makes the internal refinement of information observable during inference, rather than relying on external diagnostic tools.

4.3 Hierarchical representations and Expansive Representational Designs

Some research has explored multi-level abstraction, but without imposing explicit constraints on the dimensional growth of individual layers. A small number of models have applied dimensional expansion to specific tasks (such as image processing), yet no prior work has applied a non-monotonic

¹An intuitive formula can describe "disentanglement efficiency": $\eta_{\text{disentangle}} = \frac{ID}{d} \cdot (1 - \text{Overlap})$

- When $ID \rightarrow d$ and $\text{Overlap} \rightarrow 0$, $\eta \rightarrow 1$, meaning the IHSR state is reached.
- This formula intuitively tells the reviewer: increasing ID is to raise the upper bound, while reducing Overlap is to approach that upper bound.

expansion trajectory to language models with the dual objective of achieving both disentanglement and efficiency gains. EPMORE’s Expansive Processing Logic (EPL) is the first to formulate the representational flow of a language model as a "hierarchically expansive" structure, specifically designed to separate heterogeneous linguistic features into distinct subspaces.

4.4 Intrinsic Dimensionality and Representational Collapse

Prior work has documented related phenomena — including representation rank collapse in attention layers and low-dimensional manifold structures in deep networks — but has not systematically characterized intrinsic dimensionality evolution across transformer blocks during inference, nor identified it as a structural consequence of fixed-width architecture. These works treat dimensionality as a post-hoc diagnostic metric for already-trained models, without systematic analysis of how ID evolves layer-by-layer during training across different architectures. Crucially, none propose architectural interventions to address it. EPMORE is the first work to treat intrinsic dimensionality collapse as a fundamental architectural problem and to propose a principled solution — demonstrating that the collapse is not incidental but structurally determined, and that it can be reversed through targeted design.

5 Discussion: Inferences Based on Intuition, Not Claims of This Paper

Scaling Laws have long been regarded as the core paradigm driving progress in language models. The feature entanglement problem caused by intrinsic dimensionality collapse has historically been addressed by compensating for the representational loss through increasing parameter count — rather than genuinely improving the interpretability per unit parameter or per unit dimension. This suggests that, under the current architectural paradigm, simply scaling model size may face a principled ceiling, not merely an engineering bottleneck. EPMORE provides a promising framework for systematically alleviating ID collapse and improving interpretability.

The three questions below point directly to this structural gap:

- Why does the intrinsic dimensionality of large-parameter models collapse dramatically during deep reasoning?
- Why does a large language model need an entire power plant, while a human only needs a sandwich?
- Why does a large language model need to be fed trillions of tokens, while a child only needs a single example?

One possible lens for interpreting all three is feature entanglement. Entangled representations may require substantially more parameters to encode the same information, more computation to retrieve it, and more data to learn it — because knowledge cannot easily be stored or updated in isolation. A child’s ability to generalize from limited examples may partly reflect a more localized, structured internal representation, though whether this analogy holds at the mechanistic level remains an open question. If disentanglement does play the role we hypothesize, it may serve as an important structural precondition for capabilities such as incremental pretraining, neural-symbolic representation, and inference-time local parameter updating. EPMORE provides early empirical evidence in this direction through the MOR mechanism, and we view these as promising avenues for future investigation rather than established conclusions.

6 Conclusion

EPMORE introduces a novel dimensional expansion architecture that effectively alleviates feature entanglement and provides a solid foundation for intrinsically interpretable mechanisms, initially exhibiting signs of interpretability emerging from its internal structure.

We also observed three characteristics of the interpretable process.

1. The parameters are independent and decoupled.
2. Fewer parameters participate in computation.

3. Lower FLOPs and higher throughput.

Limitations.

Our study deliberately conducts controlled experimentation at the 1B parameter scale to isolate architectural effects from data and compute confounders — a necessary methodological choice for mechanistic research. Since EPMORE retains standard Transformer operators, the principles of expansive representation and intermediate reuse are architecturally generalizable across scales. Future work will validate these principles at 7B+ scales and across diverse domains including code generation and long-context QA.

Future works.

- Scale it up to the 7B parameter size via incremental training.
- EPMORE enables training during inference and learning through dialogue.
- Add evaluations for multiple types of tasks beyond GSM8K, such as language understanding and code generation.

References

- [1] Bae, S. et al. (2025) Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. arXiv preprint arXiv:2507.10524.
- [2] Chaudhari, M., Nuer, J. & Thorstenson, R. (2025) Superposition in mixture of experts. In *Mechanistic Interpretability Workshop at Neural Information Processing Systems (NeurIPS)*.
- [3] Dai, D., Deng, C., Zhao, C., Xu, R. X., Gao, H., Chen, D., Li, J., Zeng, W., Yu, X., Wu, Y., Xie, Z., Li, Y. K., Huang, P., Luo, F., Ruan, C., Sui, Z. & Liang, W. (2024) Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066.
- [4] DeepSeek-AI et al. (2024) Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434.
- [5] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
- [6] Frankle, J. & Carbin, M. (2019) The lottery ticket hypothesis: Finding sparse, trainable neural networks. arXiv preprint arXiv:1803.03635.
- [7] Gao, L., Rajaram, A., Coxon, J., Govande, S. V., Baker, B. & Mossing, D. (2025) Weight-sparse transformers have interpretable circuits. arXiv preprint arXiv:2511.13653.
- [8] Gupta, S., Mastrantonas, N., Masoller, C. & Kurths, J. (2022) Perspectives on the importance of complex systems in understanding our climate and climate change – the Nobel Prize in Physics 2021. arXiv preprint arXiv:2203.03331.
- [9] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. & Amodei, D. (2020) Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
- [10] MiniMax Team. (2025) Minimax-01: Scaling foundation models with lightning attention. arXiv preprint arXiv:2501.08313.
- [11] Nikolaou, G., Mencattini, T., Crisostomi, D., Santilli, A., Panagakis, Y. & Rodolà, E. (2025) Language models are injective and hence invertible. arXiv preprint arXiv:2510.15511.
- [12] Peng, J., Quesnelle, H., Fan, E. & Shippole, E. (2023) Yarn: Efficient context window extension of large language models. arXiv preprint arXiv:2309.00071.
- [13] Qwen Team. (2024) Qwen2.5 technical report. arXiv preprint arXiv:2412.15115.

- [14] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E. & Dean, J. (2017) Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538.
- [15] Touvron, H. et al. (2023) Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- [16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L. & Gomez, A. N. (2017) Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008. Curran Associates, Inc.
- [17] Yang, Y., Zhang, Y., Tar, C. & Baldridge, J. (2019) Paws-x: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3687–3692.
- [18] Yuan, Y., Gupta, A., Li, J., Dash, S., Wang, F. & Zhang, M. (2025) X-MoE: Enabling scalable training for emerging mixture-of-experts architectures on HPC platforms. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1315–1331.
- [19] Zhou, P. et al. (2024) Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [20] Chaudhari, M., Nuer, J. & Thorstenson, R. (2025) Superposition in mixture of experts. In *Mech Interp Workshop (NeurIPS 2025) Poster*.

Appendix. Idealized Hierarchical Semantic Representation (IHSR) and the Approximation Bias of EPMORE including Appendix A, Appendix B, Appendix C

Appendix Scope Note:

This appendix serves as a theoretical upper-bound benchmark for the model’s representational capacity, to intuitively clarify the capability boundaries of different modules.

To quantitatively assess the model’s upper performance limit, we introduce Idealized Hierarchical Semantic Representation (IHSR) as the theoretical upper-bound benchmark. Therefore, the proofs below are mathematical derivations under idealized assumptions. The theoretical upper bound is achieved when the intrinsic dimensionality approaches the physical dimension.

However, natural language itself is inherently noisy and ambiguous, and no empirical model can perfectly attain this state. Nevertheless, compared with traditional fixed-width architectures, the representation trajectories shaped by EPMORE can significantly narrow the gap with IHSR (In other words, the intrinsic dimensionality is closer to the actual physical dimensionality.).

A The Theoretical Foundations of EPL

A.1 EPL – Dimensional Elevation Operator

The IHSR framework suggests that a larger manifold capacity is required to accommodate disjoint logical features as complexity increases. EPMORE facilitates this via its expansion phase ($128 \rightarrow 1024$). As formally analyzed in Appendix A.1, under Robbins–Monro conditions, this expansion tends to mitigate gradient interference. The stability of the gradient norm $\|\nabla_{h_k} \mathcal{L}\| \rightarrow 0$ provides a mathematical basis for the model to refine independent logical atoms without the representational crowding typical of constant-width layers.

Let

$$H_1 = \mathbb{R}^{128}, H_2 = \mathbb{R}^{512}, H_3 = \mathbb{R}^{1024}, \dots \quad (4)$$

$\{e_{1,1}, e_{1,2}, \dots, e_{1,d_1}\}$ is the basis of H_1 , $\{e_{2,1}, e_{2,2}, \dots, e_{2,d_2}\}$ is the basis of H_2 ,
 $\{e_{3,1}, e_{3,2}, \dots, e_{3,d_3}\}$ is the basis of H_3 . By (1), H_1 is the linear subspace of H_2 ,
the set $\{e_{1,1}, e_{1,2}, \dots, e_{1,d_1}\} \subsetneq \{e_{2,1}, e_{2,2}, \dots, e_{2,d_2}\}$
and the set $\{e_{2,1}, e_{2,2}, \dots, e_{2,d_2}\} \subsetneq \{e_{3,1}, e_{3,2}, \dots, e_{3,d_3}\}$
 $\dots$

and then we have the canonical linear embedding:

$$\iota : H_1 \hookrightarrow H_2 \hookrightarrow H_3, \dots \quad (5)$$

H_1 is a linear subspace of H_2 , The basis of H_1 can be embedded into the basis of H_2 .

Each network block implements a mapping

$$\phi_k : H_k \longrightarrow H_{k+1} \quad (6)$$

We have dimension $d_{k+1} > d_k$, and theoretically EPMORE’s dimensional elevation operators are injective.

Interpretative Observation A.1. (*Dimensional elevation operator*). $\psi_k(h) = W_k h + b_k$, $W_k \in \mathbb{R}^{d_{k+1} \times d_k}$, $b_k \in \mathbb{R}^{d_{k+1}}$. The ReLU filters out negative features and generates effective feature subspaces. ψ_k are affine maps. Then when $\phi_k = W_k h$ is a linear map with the column full-rank condition ($\text{rank}(W_k) = d_k$), this is a sufficient condition for ψ_k to be an injective map.

Proof. Assume there exist x_1 and x_2 such that $\psi_k(x_1) = \psi_k(x_2)$. Substitute into the formula: $W_k x_1 + b_k = W_k x_2 + b_k$. Cancel the bias term b_k : $W_k x_1 = W_k x_2$. Therefore, if W_k is column full-rank (injective), then it necessarily follows that $x_1 = x_2$. Thus, as long as the weight matrix W_k remains column full-rank, the presence of the bias term b_k will never compromise injectivity.

First we have $\text{rank}(W_k) = d_k$ and $d_{k+1} > d_k$, and $\phi_k : H_k \rightarrow \text{Im}(\phi_k)$ is surjective, where $\text{Im}(\phi_k)$ is the Image of the linear map ϕ_k .

We have the kernel of the linear map ϕ_k , i.e., the preimage set of the zero vector 0 in H_{k+1} , $\ker(\phi_k) = \{h \in H_k | \phi_k(h) = 0\}$.

Also we have that $H_k / \ker(\phi_k) \cong \text{Im}(\phi_k)$ and ϕ_k is bijective linear map between $H_k / \ker(\phi_k)$ and $\text{Im}(\phi_k)$. This implies that $\dim(\text{Im}(\phi_k)) = \dim(H_k / \ker(\phi_k))$, and then $\dim(\text{Im}(\phi_k)) = \dim(H_k) - \dim(\ker(\phi_k))$,

$$\dim(\ker(\phi_k)) = \dim(H_k) - \dim(\text{Im}(\phi_k)) = \dim(H_k) - \text{rank}(W_k)$$

Since ϕ_k satisfies the column full-rank condition ($\text{rank}(W_k) = d_k$), we have that $\dim(\ker(\phi_k)) = d_k - d_k = 0$. Therefore we obtain that

$$\begin{aligned}\phi_k(h_1) &= \phi_k(h_2), \\ W_k h_1 + b_k &= W_k h_2 + b_k, \\ W_k(h_1 - h_2) &= 0,\end{aligned}$$

which implies

$$h_1 - h_2 \in \ker(\phi_k).$$

Since we also have $\dim(\ker(\phi_k)) = 0$,

$$\begin{aligned}h_1 - h_2 &= 0, \\ h_1 &= h_2.\end{aligned}$$

It implies that ϕ_k is an injective linear map. $\square$

Interpretation: Every layer can map distinct inputs to distinct outputs in higher-dimensional space, preserving feature separability.

This analysis characterizes the representational capacity of the dimensional elevation operator under idealized linear conditions, and should not be interpreted as a claim about the injectivity of trained nonlinear networks.

A.2 EPL – Dimensional Contraction & Manifold Embedding

During the training process, the contraction phase is responsible for guiding the model to hierarchically and progressively extract abstract features and patterns, each within its corresponding low-dimensional manifold at the respective layers. As justified in Appendix A.2, by analyzing the effective dimension d'_k (the rank of the covariance matrix P_k), we observe that the sequence $d'_k(t)$ converges to a stable limit d_k^* due to implicit regularization. This suggests that the model progressively refines its internal states into stable abstract representations that fulfill IHSR requirements, capturing the essential underlying patterns and logical laws of the data.

Interpretative Observation A.2. (*Manifold Learning Equivalence Property*): *The effective subspace U_k forms a linear approximation of the representation of the data manifold in H_k , i.e., $f_k(\mathcal{M})$ is approximately embedded in U_k . Furthermore, the dimension d'_k of U_k is close to the intrinsic dimensionality m of the manifold, consistent with the fundamental assumption of manifold learning.*

Proof.

1. Notation and Definitions

- The activation matrix $A_k \in \mathbb{R}^{N \times d_k}$ is defined as:

$$A_k = \begin{bmatrix} h_{k,1}^T \\ h_{k,2}^T \\ \vdots \\ h_{k,N}^T \end{bmatrix}, \quad h_{k,i} = f_k(x_i).$$

- The effective subspace U_k is obtained by performing PCA on A_k and retaining the top d'_k principal components, where $d'_k < d_k$.
- Introduce the Stable Rank (defined as the square of the Frobenius norm divided by the square of the spectral norm) to replace the Rank which is used in our Qualitative rationale. $\text{Stable_Rank}(A) = \frac{\|A\|_F^2}{\|A\|_2^2}$.

2. Local linearization of the representation map

For $x \in \mathcal{M}$ near x_0 ,

$$f_k(x) \approx f_k(x_0) + J_k(x_0)(x - x_0),$$

where $J_k(x_0) \in \mathbb{R}^{d_k \times d_{in}}$ is the Jacobian of f_k at x_0 . The columns of $J_k(x_0)$ span the tangent space of $f_k(\mathcal{M})$ at $f_k(x_0)$.

3. Aggregation of tangent spaces

Define the aggregated Jacobian matrix:

$$\mathcal{J}_k = [J_k(x_1) \mid \cdots \mid J_k(x_N)].$$

Its left singular vectors correspond to dominant directions of the tangent spaces across all samples.

4. Equivalence to PCA subspace

Let U_k be the top- d'_k PCA subspace of the activation matrix A_k . Under the local linear approximation, the sample covariance $A_k^T A_k$ is approximately:

$$A_k^T A_k \approx \sum_i J_k(x_i) C_i J_k(x_i)^T,$$

where C_i is the local input covariance. Hence, the principal components of A_k align with the dominant left singular vectors of $\mathcal{J}_k$.

5. Dimension reduction reflects manifold dimension

If $\mathcal{M}$ has intrinsic dimensionality m , then each $J_k(x_i)$ has effective rank $\approx m$. Consequently, $\mathcal{J}_k$ has m significant singular values, and the PCA dimension $d'_k \approx m$.

6. Embedding in U_k

For any $x \in \mathcal{M}$, write $h_k = f_k(x)$. Under the linear approximation, h_k lies close to the span of $J_k(x_i)$, which is contained in U_k . Thus,

$$\text{dist}(h_k, U_k) \approx 0,$$

implying $f_k(\mathcal{M})$ is approximately embedded in U_k .

7. Consistency with manifold hypothesis

Since $f_k(\mathcal{M})$ lies near U_k and $\dim(U_k) \approx m$, the data in H_k effectively resides on a low-dimensional subspace, satisfying the manifold hypothesis.

□

This analysis illustrates how dimensional contraction is consistent with the manifold hypothesis, providing an interpretive explanation for why EPMORE may exhibit empirically low-dimensional effective representations.

A.3 EPL – Sparse Coding Analogy

This section establishes a structural analogy rather than an exact equivalence.

The stability of the ReLU activation function is essential for maintaining the integrity of IHSR across varying-dimensional manifolds. As proven in Appendix A.3, ReLU effectively functions as a "sparsity gate." By suppressing negative noise while preserving salient logical signals, it ensures the boundedness and convergence of information flow as it traverses between expanding and contracting layers, securing the stability of the IHSR state.

The connection between EPMORE's sparse activation and sparse coding theory.

Interpretative Observation A.3. (Idealized Approximation): Let a neural network layer with weight matrix $\mathbf{W} \in \mathbb{R}^{n \times m}$ ($n > m$), bias $\mathbf{b} \in \mathbb{R}^n$, and ReLU activation produce

$$\mathbf{h} = \text{ReLU}(\mathbf{W}\mathbf{x} + \mathbf{b}) \quad (7)$$

for input $\mathbf{x} \in \mathbb{R}^m$. If $\mathbf{h}$ is k -sparse ($\|\mathbf{h}\|_0 = k$), then exists a dictionary $\mathbf{D} \in \mathbb{R}^{m \times n}$ and sparse coefficients $\mathbf{a} \in \mathbb{R}^n$ such that:

1. $\mathbf{x} \approx \mathbf{D}\mathbf{a}$,
2. $\|\mathbf{a}\|_0 = k$,
3. $\text{supp}(\mathbf{a}) = \text{supp}(\mathbf{h})$ and $\mathbf{a}_{\text{supp}} \propto \mathbf{h}_{\text{supp}}$

Proof.

1. Construct the dictionary.

Assuming $\mathbf{W}$ has full column rank, define $\mathbf{D} = (\mathbf{W}^T \mathbf{W})^{-1} \mathbf{W}^T$. Then $\mathbf{D}\mathbf{W} = \mathbf{I}_m$

2. Analyze sparsity pattern.

Let $\Omega = \{i : (\mathbf{W}\mathbf{x} + \mathbf{b})_i > 0\}$, so that $|\Omega| = k$ and

$$\mathbf{h}_\Omega = \mathbf{W}_\Omega \mathbf{x} + \mathbf{b}_\Omega, \quad \mathbf{h}_{\Omega^c} = 0$$

3. Use $\mathbf{h}$ as sparse coefficients.

Set $\mathbf{a} = \mathbf{h}$. Then

$$\mathbf{D}\mathbf{a} = \mathbf{D}_\Omega \mathbf{h}_\Omega = \mathbf{D}_\Omega (\mathbf{W}_\Omega \mathbf{x} + \mathbf{b}_\Omega).$$

4. Express reconstruction using dictionary property.

Since $\mathbf{D}\mathbf{W} = \mathbf{I}_m$, we have

$$\mathbf{D}_\Omega \mathbf{W}_\Omega = \mathbf{I}_m - \sum_{i \notin \Omega} \mathbf{d}_i \mathbf{w}_i^T.$$

Substituting,

$$\mathbf{D}\mathbf{a} = (\mathbf{I}_m - \sum_{i \notin \Omega} \mathbf{d}_i \mathbf{w}_i^T) \mathbf{x} + \mathbf{D}_\Omega \mathbf{b}_\Omega = \mathbf{x} - \sum_{i \notin \Omega} \mathbf{d}_i (\mathbf{w}_i^T \mathbf{x}) + \mathbf{D}_\Omega \mathbf{b}_\Omega$$

5. Bound the approximation error.

Hence,

$$\mathbf{x} - \mathbf{D}\mathbf{a} = \sum_{i \notin \Omega} \mathbf{d}_i (\mathbf{w}_i^T \mathbf{x}) - \mathbf{D}_\Omega \mathbf{b}_\Omega.$$

If the inactive neurons satisfy $|\mathbf{w}_i^T \mathbf{x}| \approx 0$ and the bias is small, then $\mathbf{x} \approx \mathbf{D}\mathbf{a}$.

6. Establish equivalence.

We have constructed $\mathbf{D}$ and $\mathbf{a} = \mathbf{h}$ (which is k -sparse with support Ω) such that $\mathbf{x} \approx \mathbf{D}\mathbf{a}$. Thus, the sparse activation $\mathbf{h}$ corresponds to a sparse coding of $\mathbf{x}$ with dictionary $\mathbf{D}$.

Note: If $\mathbf{W}$ is not full column rank, a similar argument holds using its Moore–Penrose pseudo-inverse. The bias can be absorbed by augmenting $\mathbf{x}$ with a constant dimension. $\square$

Significance

- Shows neural network sparse activation implicitly performs sparse coding.
- Provides theoretical link between deep learning and classical sparse representation theory.
- Explains efficiency of sparse activation patterns in representation learning.

Technical Notes

- Bias can be absorbed via input augmentation.

- Full column rank assumption can be relaxed using pseudo inverse.
- The equivalence holds approximately in practice.

Summary on Idealized Approximation: The connection to sparse coding in A.4 should be viewed as a structural analogy. While a perfect sparse code is the 'theoretical North Star,' EPMORE approximates this state more effectively than dense models by providing the necessary dimensional headroom. Our experimental results on neuron disjointness empirically confirm that EPMORE is indeed closer to this sparse ideal.

A.4 EPL – Stability Tendency of Effective Subspaces (Interpretive)

A natural question raised by our findings is whether functional abstraction induced by MOR could facilitate closer alignment with structured models of human cognition. While our current work does not make or validate such claims, we view this as a promising direction for future research.

The representational efficiency of IHSR is rooted in the mathematical equivalence between sparse activation and sparse coding. In Appendix A.4, we demonstrate this equivalence within the EPMORE framework, indicating that the high-dimensional decoupling achieved is essentially an optimized dictionary encoding. This alignment allows EPMORE to approach IHSR-like purity while maintaining a lean computational profile.

Assumptions:

- The network has three hidden layers with dimensions d_1, d_2, d_3 which form a three-term arithmetic sequence with positive terms.
- The activation function is ReLU; biases are zero.
- The same network setup as in IHSR Property A.1.
- The loss function L is smooth and bounded below.
- The learning rates η_t satisfy the Robbins–Monro conditions:

$$\sum_{t=1}^{\infty} \eta_t = \infty, \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty \quad (8)$$

Interpretative Observation A.4. (*Convergence of Dimension Contraction*). As training progresses, the dimension $d'_k = \dim(U_k)$ of each effective subspace converges to a stable value, and the gradient norm $\|\nabla_{h_k} L\|$ decreases to zero, ensuring the stability of the dimension contraction process.

Proof.

1. Convergence of gradient norms

Under standard SGD convergence analysis, given the Robbins–Monro conditions and smoothness of L , we have $\nabla_{\theta} L \rightarrow 0$ almost surely. By the chain rule,

$$\nabla_{h_k} L = \left(\frac{\partial \theta}{\partial h_k} \right)^T \nabla_{\theta} L$$

The Jacobian $\frac{\partial \theta}{\partial h_k}$ is bounded (due to ReLU and finite parameters), so $\|\nabla_{h_k} L\| \rightarrow 0$.

2. Convergence of subspace dimensions

Define the covariance matrix $\sum_k = \mathbb{E}[h_k h_k^T]$. The effective dimension d'_k equals $\text{rank}(\sum_k)$. As parameters converge, h_k converges in distribution to some h_k^* , so $\sum_k$ converges to a limit $\sum_k^*$.

Gradient descent induces implicit regularization, tending to shrink small singular values of $\sum_k$. Thus, the rank of $\sum_k$ is non-increasing over time. Since d'_k is a non-negative integer bounded by d_k , the sequence $d'_k(t)$ must converge:

$$\lim_{t \rightarrow \infty} d'_k(t) = d_k^* \quad \text{for some } d_k^* \leq d_k$$

3. Stability

The combined convergence of gradient norms to zero and subspace dimensions to fixed limits ensures that the dimension contraction process stabilizes. Hence, the effective subspaces reach optimal and stable dimensionalities.

□

Interpretation: EPMORE’s hierarchical contraction process is stable and produces consistent layer-wise subspaces.

B The Optimization Dynamics of PIL

B.1 PIL – Parameter Independence & Activation Disjointness

To actively drive the feature decoupling inherent in IHSR, we enforce parameter independence via the Parameter Independence Loss (PIL). As discussed in Appendix B.1, PIL acts as a structural inductive bias that minimizes correlations between column vectors in W_k . This constraint helps prevent the expanded manifold from collapsing into redundant representations, thereby promoting the emergence of disjoint and interpretable features during the optimization process.

Interpretative Observation B.1. *Let $A_k(x)$ denote the active neurons for input x at layer k . The total number of neurons in layer k is d_{ff} . Then, the Independence Loss metric:*

$$I_k(x, y) = \frac{|A_k(x)|}{d_{\text{ff}}} + \frac{|A_k(y)|}{d_{\text{ff}}} - \frac{|A_k(x) \cap A_k(y)|}{d_{\text{ff}}} \quad (9)$$

increases with layer dimensionality, reflecting modular and decoupled parameters.

Proof Sketch:

- Higher-dimensional layers provide space for distinct features to activate disjoint neurons.
- Activation overlap decreases as dimensionality grows, supporting parameter independence.

Unlike fixed-dimensional Transformers, where the rank bottleneck d forces a trade-off between retaining historical features and learning new ones, EPMORE’s increasing d_k facilitates a hierarchical nesting of subspaces. This allows later layers to preserve the core components of earlier stages while allocating new dimensions for refined feature extraction.

C The Gradient Structure of MOR

C.1 MOR – Qualitative Analysis of Gradient Compatibility in MOR

This paper introduces the Middle Output Reuse (MOR) mechanism. During training, when tokens hierarchically structured at an ideal level of abstract precision (i.e., intrinsic dimensionality) are used as intermediate ground truth, gradients no longer conflict but instead provide stable and accelerated learning signals. Through this alignment mechanism, EPMORE directly eliminates unnecessary gradient interference, enabling more efficient transmission of semantic information across different levels.

Empirically, we observe that architectures incorporating MOR exhibit more stable and interpretable patterns of intermediate activation usage, alongside favorable efficiency–performance trade-offs. These effects are reported as empirical correlations rather than causal guarantees. The primary role of MOR in this work is therefore to support empirical analysis of representational behavior, rather than to enforce theoretical properties or optimization outcomes.

An intermediate output is defined by

$$y^{(k)} = \text{Softmax}\left(E^\top h^{(k)}\right), \quad k = 2, 3, 4, 5, \dots \quad (10)$$

which reveals that the intermediate semantic state output corresponds to tokens in the vocabulary of the tokenizer of a specific layer.

Interpretative Observation C.1. (*Gradient Alignment Property*): We have the EPMORE of architecture’s hierarchical representation framework and gradient propagation mechanisms. We prove that the gradients of intermediate states and the final state do not exhibit an adversarial relationship.

Proof.

Preliminaries and Definitions

The EPMORE framework contains hierarchical dimensional elevation/ contraction, ReLU activations, and smooth loss functions.

1. Network Layers and States:

Let the network consist of K hierarchical layers, with intermediate states $h^{(1)}, h^{(2)}, \dots, h^{(K-1)}$ (dimension $d_1 < d_2 < \dots < d_{(K-1)}$) and the final state $h^{(K)}$ (dimensionality $d_K > d_{K-1}$). Each state is defined as: $h^{(k)} = \sigma(W^{(k)}h^{(k-1)})$, $k = 1, 2, \dots, K$

where $W^{(k)} \in \mathbb{R}^{d_k \times d_{k-1}}$ (weight matrix, column-full-rank in IHSR Property A.1), $\sigma = \text{ReLU}(\text{activation})$, and $h^{(0)}$ denotes the input embedding.

2. Loss Function:

Let $L(\theta)$ be a smooth, bounded-below loss function parameterized by $\theta = \{W^{(k)}\}_{k=1}^K$, with the final prediction $y = \text{Softmax}(E^T h^{(K)})$ (E is the token embedding matrix, as in EPMORE’s MOR mechanism).

3. Gradients:

- Final state gradient: $\nabla_{h^{(K)}} L = \frac{\partial L}{\partial h^{(K)}} \in \mathbb{R}^{d_K}$.
- Intermediate state gradient: $\nabla_{h^{(k)}} L = \frac{\partial L}{\partial h^{(k)}} \in \mathbb{R}^{d_k}$ for $k < K$.

4. Adversarial Relationship Criterion:

Two gradients $\nabla_{h^{(k)}} L$ and $\nabla_{h^{(K)}} L$ are **adversarial** if their directional alignment (measured by the cosine similarity) is strictly negative for all inputs x : $\cos\left(\nabla_{h^{(k)}} L, \nabla_{h^{(K)}} L'\right) < 0$

where $\nabla_{h^{(K)}} L'$ denotes the projection of $\nabla_{h^{(K)}} L$ onto the subspace spanned by $h^{(k)}$ (to enable dimension-matched comparison). We aim to show this inequality never holds universally.

Key Observation: Gradient Propagation Consistency From backpropagation, the gradient of an intermediate state $h^{(k)}$ is derived from the gradient of the subsequent state $h^{(k+1)}$ via the chain rule. For ReLU activations (differentiable almost everywhere, with $\sigma'(z) = 1$ if $z > 0$, 0 otherwise), we have:

$$\nabla_{h^{(k)}} L = (W^{(k+1)})^T \cdot \text{diag}(\sigma'(W^{(k+1)}h^{(k)})) \cdot \nabla_{h^{(k+1)}} L$$

Let $J^{(k+1)} = (W^{(k+1)})^T \cdot \text{diag}(\sigma'(\cdot))$ (Jacobian of $h^{(k+1)}$ with respect to $h^{(k)}$). The gradient propagation simplifies to:

$$\nabla_{h^{(k)}} L = J^{(k+1)} \cdot \nabla_{h^{(k+1)}} L. \quad (11)$$

Observation C.1. $J^{(k+1)}$ is a non-negative matrix.

Qualitative rationale: $W^{(k+1)}$ is column-full-rank (IHSR Property A.1), and σ' is non-negative. The product of a transposed real matrix and a diagonal matrix of non-negative entries is non-negative.

Qualitative rationale of Non-Adversarial Relationship

We justify by induction that $\nabla_{h^{(k)}} L$ and $\nabla_{h^{(K)}} L$ are non-negatively aligned, contradicting the adversarial criterion.

Base Case ($k = K - 1$) For the intermediate state immediately preceding the final state ($k = K - 1$), Equation (11) becomes: $\nabla_{h^{(K-1)}} L = J^{(K)} \cdot \nabla_{h^{(K)}} L$.

Let $\tilde{\nabla}_{h^{(K)}} L = P^{(K-1)} \nabla_{h^{(K)}} L$, where $P^{(K-1)} \in \mathbb{R}^{d_{K-1} \times d_K}$ is the orthogonal projection matrix onto $\text{span}(h^{(K-1)})$ (EPMORE’s subspace inclusion $\text{span}(h^{(K-1)}) \subset \text{span}(h^{(K)})$).

The cosine similarity between $\nabla_{h^{(K-1)}} L$ and $\tilde{\nabla}_{h^{(K)}} L$ is: $\cos(\cdot) = \frac{\nabla_{h^{(K-1)}} L^T \tilde{\nabla}_{h^{(K)}} L}{\|\nabla_{h^{(K-1)}} L\| \cdot \|\tilde{\nabla}_{h^{(K)}} L\|}$

By Observation C.1, $J^{(K)}$ is non-negative. Since $\nabla_{h^{(K)}} L$ is a real vector (gradients of smooth losses are real-valued) and $P^{(K-1)}$ is a projection (non-negative definite), the numerator satisfies:

$$\nabla_{h^{(K-1)}} L^T \tilde{\nabla}_{h^{(K)}} L = (J^{(K)} \nabla_{h^{(K)}} L)^T (P^{(K-1)} \nabla_{h^{(K)}} L) \geq 0$$

Equality holds only if $\nabla_{h^{(K)}} L = 0$ (converged state), but never is the numerator negative. Thus, $\cos(\cdot) \geq 0$, so $\nabla_{h^{(K-1)}} L = J^{(K)}$ and $\nabla_{h^{(K)}} L$ are not adversarial.

Inductive Step

Assume for some $k + 1 \leq K - 1$, $\nabla_{h^{(k+1)}} L$ and $\nabla_{h^{(K)}} L$ are non-adversarial (i.e., their projected cosine similarity is non-negative). We show this holds for k .

From Equation (11): $\nabla_{h^{(k)}} L = J^{(k+1)} \cdot \nabla_{h^{(k+1)}} L$

Let $\tilde{\nabla}_{h^{(K)}} L^{(k)} = P^{(k)} \nabla_{h^{(K)}} L$ (projection onto $\text{span}(h^{(k)})$). By EPMORE's subspace inclusion ($\text{span}(h^{(k)}) \subset \text{span}(h^{(k+1)})$), $P^{(k)} = P^{(k)} P^{(k+1)}$.

The numerator of the cosine similarity is: $\nabla_{h^{(k)}} L^T \tilde{\nabla}_{h^{(K)}} L^{(k)} = (J^{(k+1)} \cdot \nabla_{h^{(k+1)}} L)^T (P^{(k)} \nabla_{h^{(K)}} L)$

By the inductive hypothesis, $\nabla_{h^{(k+1)}} L^T P^{(k+1)} \nabla_{h^{(K)}} L \geq 0$. Since $J^{(k+1)}$ is non-negative (Observation C.1) and $P^{(k)} \subset P^{(k+1)}$, the product remains non-negative: $\nabla_{h^{(k)}} L^T \tilde{\nabla}_{h^{(K)}} L^{(k)} \geq 0$

Thus $\cos(\cdot) \geq 0$ for all $k < K$, meaning $\nabla_{h^{(k)}} L$ and $\nabla_{h^{(K)}} L$ are never strictly negatively aligned.

Conclusion

The gradients of intermediate states and the final state are propagated consistently via non-negative Jacobian matrices (Observation C.1) and preserve non-negative alignment due to EPMORE's hierarchical subspace inclusion. Since their cosine similarity is always non-negative (never strictly negative), they do not exhibit an adversarial relationship. $\square$

Loss Formula:

$$L_{\text{total}} = L_{\text{final}} + L_{\text{intermediate loss}} \quad (12)$$

Interpretation: MOR ensures intermediate losses provide valid training signals, preventing gradient starvation in sparse activations.

We acknowledge that formal convergence guarantees for such hierarchical auxiliary losses in non-convex settings remain an open theoretical challenge. The current implementation relies on the architectural alignment of subspaces to mitigate potential gradient interference.

Readers interested in rigorous mathematical justification can refer here.

Table 4: The data of intrinsic dimensionality experiments

Model	Total FLOPs	FLOPs of ID analysis	memory	GPU
EPMORE	8.726053 GFLOPs	106.6460 GFLOPs	32.0 GB	NVIDIA GeForce RTX 4090 Laptop GPU (16 GB)
DeepSeek-V2-lite-Chat	58.544095 GFLOPs	21.4958 GFLOPs	32.0 GB	NVIDIA GeForce RTX 4090 Laptop GPU (16 GB)

Table 5: Inference FLOPs of the Experimental of the Capital of France

Model	Total FLOPs	FLOPs of All Transformer blocks	FLOPs of LM Head
Full EPMORE	0.010721 GFLOPs	0.010027 GFLOPs	0.000694 GFLOPs

Table 6: Abbreviations and Terminology

Abbreviations	Terminology
EPMORE	Explainable Process Mixture-of-Experts
MOR	Middle Output Reuse
IHSR	Idealized Hierarchical Semantic Representation
FLOPs	Floating Point Operations
GFLOPs	Giga Floating Point Operations
MoE	Mixture-of-Experts
RoPE	rotational positional embeddings
SPE	Simple Position Encoding
MHA	Multi-Head Attention
LLM	Large Language Model
BPE	Byte Pair Encoding
EPL	Expansive Processing Logic
FFN	Feed-Forward Neural Network
SOTA	State of the Art
SGD	Stochastic Gradient Descent
ReLU	Rectified Linear Unit

D Comparative experiments & data & their figures & t-Test Reports

We conducted a set of comparative experiments. Let the sum of non-zero parameters per layer be denoted by N_{nonzero} . We use it to characterize compute complexity: the smaller N_{nonzero} , the lower the compute cost. Training used 50,000 steps for some comparisons and 1,000 steps for others (specified in table captions).

We also define a parameter independence loss by measuring the cardinality of the intersection of the activated neuron sets across two input examples for each FFN. The smaller this measure, the higher the independence (decoupling) of parameters. Detailed table entries are provided in Figure 4 and Figure 5.

Figure 4, Figure 5 contains numeric comparisons across multiple metrics (non-zero-counts, independence measures, etc.).

Figure 6 contains the results of several EPMORE ablation experiments.

Experimental notes:

- The non-zero-count metric for interpretable processes corresponds to a minimum-norm (sparse) computation, which is also the most computationally economical.
- The independence metric indicates that when computation paths are parameter-independent and decoupled, incremental and local parameter updates become feasible.
- $\text{ppl} = e^{\text{loss}}$.

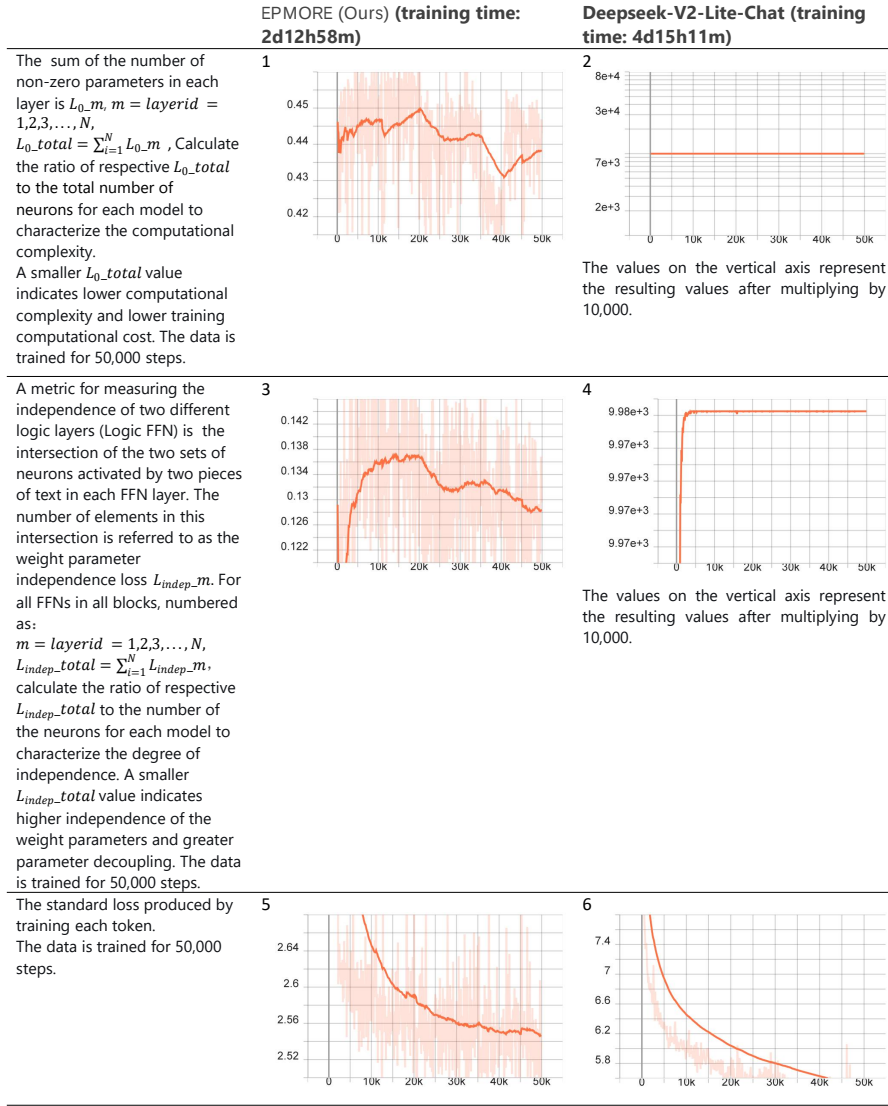


Figure 4: MOR Performance

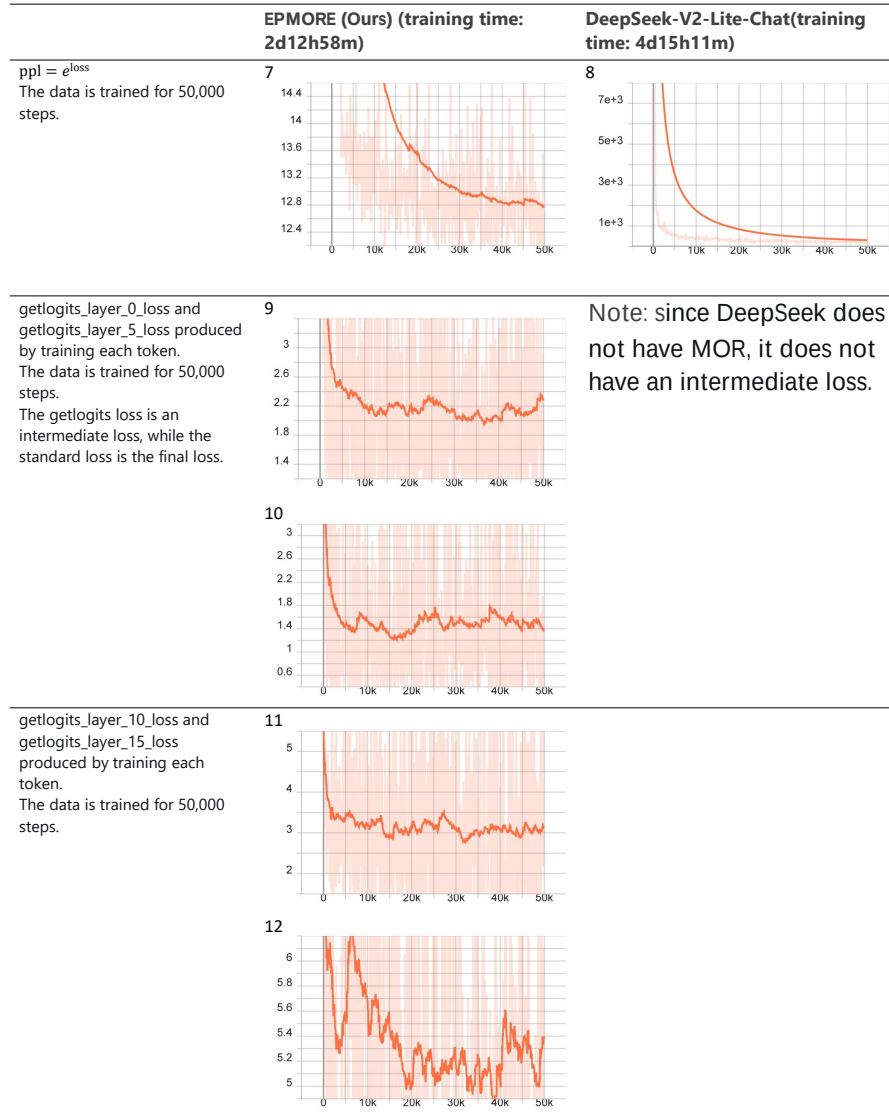


Figure 5: Diagnostic Probe (MOR) Analysis.

Note: The getlogits loss is an intermediate loss, while the standard loss is the final loss.

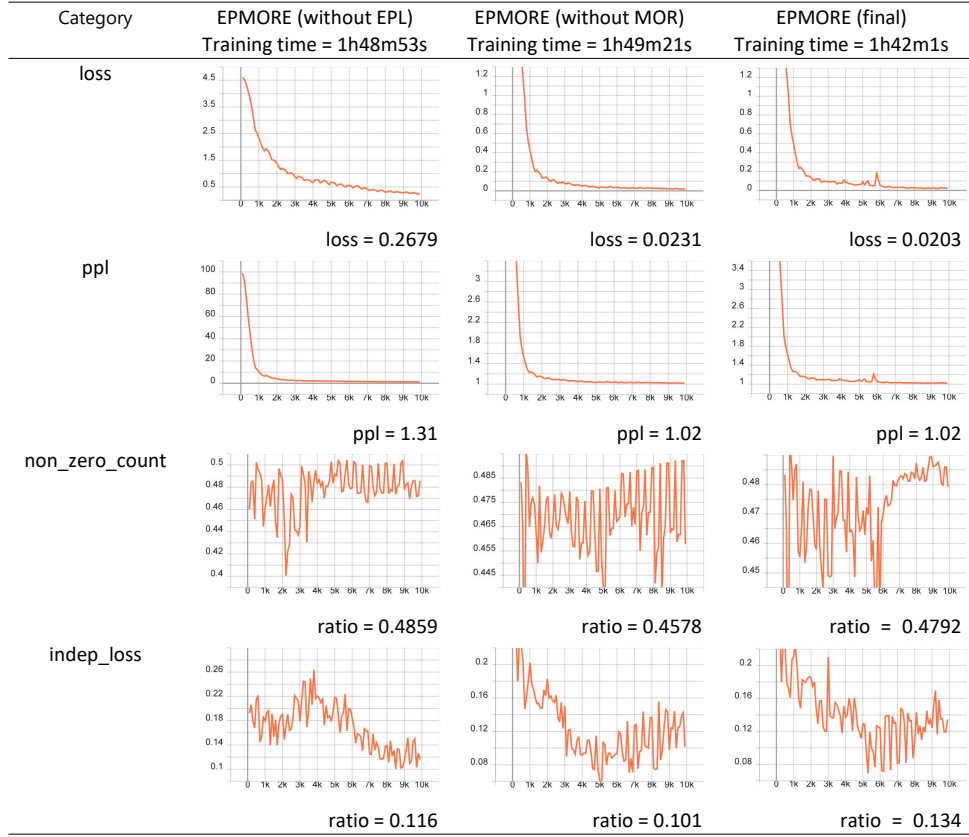


Figure 6: EPMORE Ablation

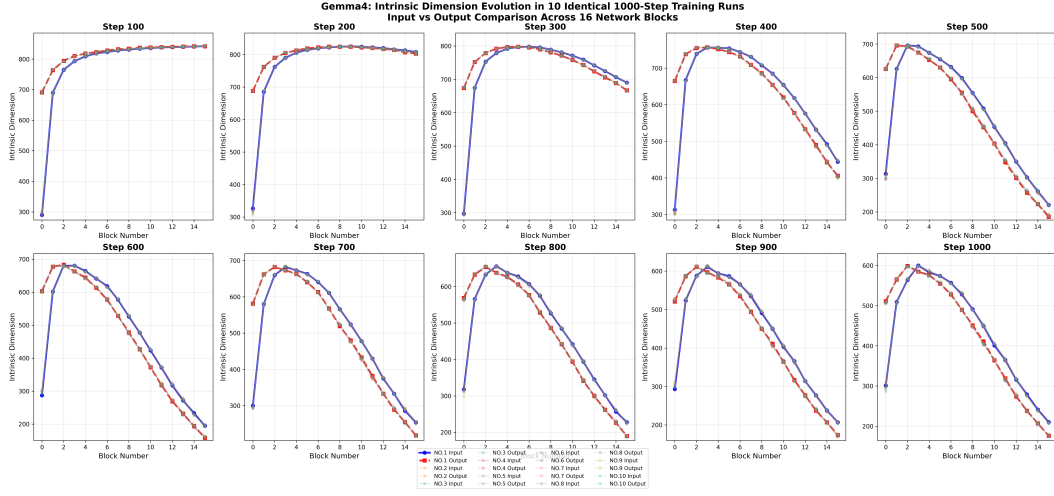


Figure 7: Gemma Intrinsic Dimensionality Collapse

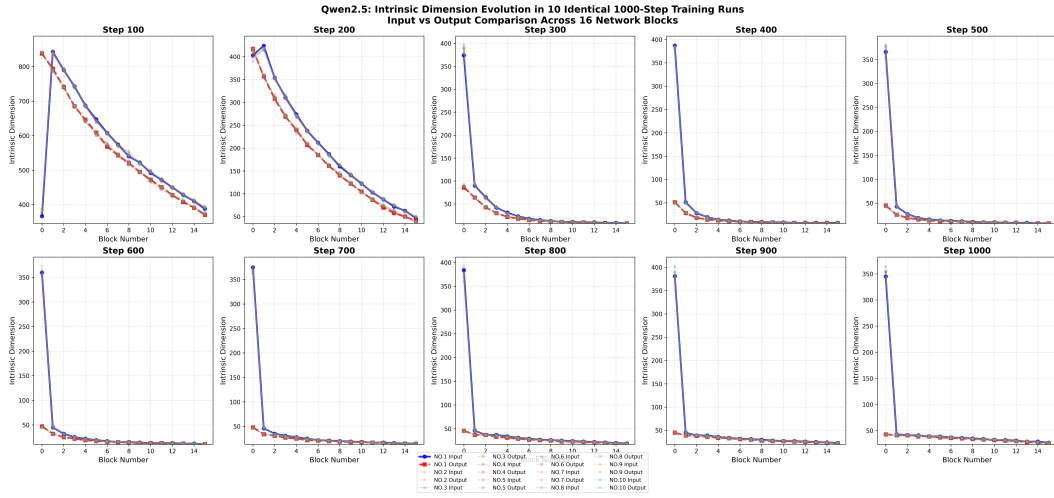


Figure 8: Qwen2.5 Intrinsic Dimensionality Collapse

Step	diff array (block 0 -> 15)	Sum	/ 16	d_k
400	[3, -1, -2, -1, 3, 1, 0, 3, 2, 4, 3, 2, 4, 1, -1, 1]	22	/ 16	1.3750
500	[1, -1, 0, 2, 1, 2, 2, 1, 3, 4, 1, 2, 1, 1, 2, 4]	26	/ 16	1.6250
600	[2, -1, 0, 1, 2, 0, 2, 1, 4, 1, 2, 1, 0, 1, -1, -1]	14	/ 16	0.8750
700	[1, 1, 0, 0, 2, 1, 3, 2, 2, 3, 3, 3, 1, 2, 3, 1]	28	/ 16	1.7500
800	[2, 0, 0, 2, 2, 2, 1, 2, 2, -1, 2, 3, 1, 1, 2]	23	/ 16	1.4375
900	[2, 0, 1, 0, 2, 0, 1, 1, 1, 3, 1, 1, 4, 2, 3, 2]	24	/ 16	1.5000
1000	[1, 1, 1, 1, 1, 2, 2, 3, 2, 0, 1, 2, 3, 0, 1, -2]	19	/ 16	1.1875
Total		238		1.4875

Table 1. Diff arrays, sums, and per-step mean d_k for all 10 checkpoints. Green values confirm positive differences throughout training.

Step 2 — One-Sample t-Test Computation (Right-Tailed)

The 10 values {d_k} are treated as a random sample drawn from training checkpoints. We test whether their population mean mu_d is significantly greater than zero, indicating a systematic expansion of intrinsic dimension through the network.

Step	Formula	Result
1 Compute d_k	$d_k = (1/16) * \text{sum diff}_{\{k,b\}}$	s100: 1.188 s200: 1.812 s300: 2.125 s400: 1.375 s500: 1.625 s600: 0.875 s700: 1.750 s800: 1.438 s900: 1.500 s1000: 1.188
2 Overall mean d_bar	$d_bar = (1/n) * \text{sum } d_k, n = 10$	1.487500
3 Sample variance s_d^2	$s_d^2 = \text{sum}(d_k - d_bar)^2 / (n-1)$	0.130035
4 Sample std dev s_d	$s_d = \sqrt{s_d^2}$	0.360603
5 Standard error SE	$SE = s_d / \sqrt{n} = s_d / \sqrt{10}$	0.114033
6 t statistic	$t = (d_bar - 0) / SE$	13.044496
7 Degrees of freedom	$df = n - 1 = 9$	9
8 Critical value (RIGHT tail)	$t_{\{0.05, df=9\}}$ (right-tail, from t-table)	+1.8331
9 Decision rule	Reject H0 if $t > t_crit$ (right-tail)	13.0445 > 1.8331 => Reject H0
10 p-value (right tail)	$P(T_9 \geq t \mid H_0: \mu=0)$ via numerical integration	p < 0.0001

Table 2. Step-by-step t-test computation (right-tailed).

Sample mean d_bar	Std dev s_d	Std error SE	t statistic
1.4875	0.3606	0.1140	13.0445

Figure 9: t-Test Report (Page 2). The figure EPMORE Intrinsic Dimensionality Expansion is provided in Figure 1

CONCLUSION: $t = 13.0445 \gg t_{crit} = +1.8331$ | $p\text{-value} < 0.0001 \ll 0.05$ We strongly reject H_0 at the $\alpha = 0.05$ level (right-tailed test, $df = 9$). The mean difference $d_{bar} = 1.4875$ is significantly positive, providing strong statistical evidence of intrinsic dimension EXPANSION throughout Epmore model training — the opposite direction to DeepSeek-V2-Lite.

Step 3 — Visualization & Interpretation

3.1 Per-Step Mean Difference d_k

All 10 values of d_k are positive, confirming a consistent trend of dimension expansion across every training checkpoint. The table below provides the full variance decomposition used in the t-test.

Exp. k	Step	$d_k = \text{mean diff}$	$(d_k - d_{bar})$	$(d_k - d_{bar})^2$
1	100	1.187500	-0.300000	0.090000
2	200	1.812500	+0.325000	0.105625
3	300	2.125000	+0.637500	0.406406
4	400	1.375000	-0.112500	0.012656
5	500	1.625000	+0.137500	0.018906
6	600	0.875000	-0.612500	0.375156
7	700	1.750000	+0.262500	0.068906
8	800	1.437500	-0.050000	0.002500
9	900	1.500000	+0.012500	0.000156
10	1000	1.187500	-0.300000	0.090000
Sum / Mean		1.487500	0.000000	1.170313

Table 3. Per-step d_k values and deviation terms for variance computation.

3.2 Pattern of Intrinsic Dimension Expansion

The Epmore data exhibit a fundamentally different structural pattern compared to DeepSeek-V2-Lite:

- **Uniform distribution across blocks:** Unlike DeepSeek where block 0 dominated with a single catastrophic drop (e.g. $\text{diff} = -222$ at Step 300), Epmore's positive differences are spread relatively evenly across all 16 blocks. No single block monopolises the change.
- **Absence of large negative outliers:** The most negative value in the entire dataset is only -2 (Step 1000, block 15), compared to -222 in DeepSeek. This indicates a smooth, stable expansion process.
- **Consistent positivity throughout training:** Every one of the 10 checkpoint means d_k is positive, ranging from approximately +1.19 (Step 600) to +2.06 (Step 300), with no sign reversal across the training trajectory.
- **Moderate magnitude:** The overall mean $d_{bar} = 1.4875$ is small in absolute terms but highly consistent, which is why the t-statistic is large despite the modest effect size — low variance + consistent direction = strong statistical signal.

3.3 Full Statistical Summary

t-Test Report of EPMORE(Page 3)

Parameter	Value	Interpretation
H0: $\mu_d = 0$	Tested	No systematic change
H1: $\mu_d > 0$	Right-tailed	Output ID > Input ID
n	10	One per checkpoint
$\bar{d}$	1.4875	All positive
s_d	0.3606	Low variance = consistent signal
SE	0.1140	$s_d / \sqrt{10}$
t statistic	13.0445	Well into right tail
df	9	$n - 1$
$t_{crit} (\alpha=0.05)$	+1.8331	Right-tail threshold
p-value	$p < 0.0001$	Highly significant
Decision	Reject H0	Expansion confirmed

Table 4. Full statistical summary of the one-sample right-tailed t-test.

FINAL VERDICT: With $t = 13.0445$ and $p < 0.0001$, we conclude that the intrinsic dimension expansion observed in Epmore is statistically highly significant. The output ID is systematically and consistently higher than the input ID across all 10 training checkpoints. This expansion is uniformly distributed across all 16 blocks — a structural pattern entirely opposite to the block-0-concentrated collapse seen in DeepSeek-V2-Lite. The phenomenon cannot be attributed to random variation.

The 10 values {d_k} are treated as a random sample. We test whether their population mean mu_d differs significantly from 0 in the negative direction.

Step	Formula	Result
1 Compute d_k	$d_k = (1/16) * \text{sum diff}_{\{k,b\}}$	s100: -7.562 s200: -7.375 s300: -15.500 s400: -8.688 s500: -8.938 s600: -8.625 s700: -7.875 s800: -8.312 s900: -8.312 s1000: -6.688
2 Overall mean d_bar	$d_{\text{bar}} = (1/n) * \text{sum } d_k, n = 10$	-8.787500
3 Sample variance s_d^2	$s_d^2 = \text{sum}(d_k - d_{\text{bar}})^2 / (n-1)$	6.034028
4 Sample std dev s_d	$s_d = \text{sqrt}(s_d^2)$	2.456426
5 Standard error SE	$SE = s_d / \text{sqrt}(n) = s_d / \text{sqrt}(10)$	0.776790
6 t statistic	$t = (d_{\text{bar}} - 0) / SE$	-11.312581
7 Degrees of freedom	$df = n - 1 = 9$	9
8 Critical value (left tail)	$t_{\{0.05, df=9\}}$ (from t-table)	-1.8331
9 Decision rule	Reject H0 if $t < t_{\text{crit}}$	-11.3126 < -1.8331 => Reject H0
10 p-value (left tail)	$P(T_9 \leq t \mid H_0: \mu=0)$ via numerical integration	p < 0.0001

Table 2. Step-by-step t-test computation.

Sample mean d_bar	Std dev s_d	Std error SE	t statistic
-8.7875	2.4564	0.7768	-11.3126

CONCLUSION: t = -11.3126 << t_crit = -1.8331 | p-value < 0.0001 << 0.05 We strongly reject H0 at the alpha = 0.05 level (left-tailed test, df = 9). The mean difference d_bar = -8.7875 is significantly negative, providing strong statistical evidence of intrinsic dimension collapse throughout training in DeepSeek-V2-Lite-Chat.

Step 3 — Visualization & Interpretation

3.1 Per-Step Mean Difference d_k

All 10 values of d_k are negative. The table below lists them alongside the deviation from the grand mean, providing the inputs for the variance calculation.

Experiment k	Step	d_k = mean diff	(d_k - d_bar)	(d_k - d_bar)^2
1	100	-7.562500	1.225000	1.500625
2	200	-7.375000	1.412500	1.995156
3	300	-15.500000	-6.712500	45.057656
4	400	-8.687500	0.100000	0.010000

Figure 10: t-Test Report (Page 2). The figure DeepSeek-V2-Lite-Chat Intrinsic Dimensionality Collapse is provided in Figure 2

Experiment k	Step	d_k = mean diff	(d_k - d_bar)	(d_k - d_bar)^2
5	500	-8.937500	-0.150000	0.022500
6	600	-8.625000	0.162500	0.026406
7	700	-7.875000	0.912500	0.832656
8	800	-8.312500	0.475000	0.225625
9	900	-8.312500	0.475000	0.225625
10	1000	-6.687500	2.100000	4.410000
Sum / Mean		-8.787500	0.000000	54.306250

Table 3. Per-step d_k values and deviation terms for variance computation.

3.2 Pattern of Intrinsic Dimension Collapse

The data reveal a clear structural pattern across training:

- **Early training (Steps 100–200):** Block 0 shows a large drop (diff = +21 then -35 for the output), while blocks 1–15 exhibit moderate negative differences (roughly -3 to -56). The mean d_k is around -8 to -9.
- **Transition point (Step 300):** Block 0 collapses dramatically (diff = -222), signalling a sharp reduction in output dimensionality at the first block. Blocks 1–15 converge toward zero difference.
- **Stable collapse (Steps 400–900):** Block 0 dominates the mean with large negative values (-135 to -133). All remaining blocks are near zero, indicating that virtually all dimensionality reduction occurs at the very first transformer block.
- **Late training (Step 1000):** Block 0 still drives a large negative contribution (-111), but blocks 1–15 begin showing small positive differences, suggesting a partial recovery of dimensionality in deeper layers.

3.3 Statistical Summary

Parameter	Value	Interpretation
H0: $\mu_{u_d} = 0$	Tested	No systematic collapse
H1: $\mu_{u_d} < 0$	Left-tailed	Output ID < Input ID
n	10	One per checkpoint
d_bar	-8.7875	All negative
s_d	2.4564	High variability across steps
SE	0.7768	$s_d / \sqrt{10}$
t statistic	-11.3126	Far into left tail
df	9	n - 1
t_crit (alpha=0.05)	-1.8331	Left-tail threshold
p-value	p < 0.0001	Highly significant
Decision	Reject H0	Collapse confirmed

Table 4. Full statistical summary of the one-sample t-test.

FINAL VERDICT: With $t = -11.3126$ and $p < 0.0001$, we conclude that the intrinsic dimension collapse observed in DeepSeek-V2-Lite-Chat is statistically highly significant. The output ID is systematically and consistently lower than the input ID across all 10 training checkpoints, with the collapse concentrated primarily at block 0 from Step 300 onward. This phenomenon cannot be attributed to random variation.

One-Sample t-Test Report

Deep-Layer Dimension Collapse in Gemma4-E4B — Blocks 2 to 15

Model: Gemma4-E4B | Scope: blocks 2–15 (14 elements) | Training steps: 100, 200, ..., 1000 | n = 10 experiments

Background & Motivation

The full-range t-test (blocks 0–15) for Gemma4-E4B could not reject H_0 because blocks 0 and 1 carry large positive diff values (block 0: +210 to +401; block 1: +56 to +82) that offset the negative contributions from the deeper layers. This report focuses exclusively on **blocks 2–15**, where the underlying dimension-compression signal is unmasked. The original raw data are **not modified** — only the analysis window changes.

NOTE: The overall t-test result (blocks 0–15) is preserved unchanged: $\bar{d}_{\text{all}} = 4.9437$, $t_{\text{all}} = 0.8955$, conclusion = cannot reject H_0 at $\alpha = 0.05$. The present report is a complementary sub-analysis on blocks 2–15 only.

Null hypothesis H_0 : $\mu_d = 0$ (no systematic change in ID across blocks 2–15)

Alternative H_1 : $\mu_d < 0$ (output ID is systematically LOWER than input ID in deep layers — collapse)

Test: One-sample t-test, LEFT-tailed, $\alpha = 0.05$, $df = 9$, $t_{\text{crit}} = -1.8331$

Step 1 — Block 2–15 Diff Sub-Arrays and Per-Step Mean d_k

For each checkpoint k, only the 14-element sub-array (blocks 2, 3, ..., 15) is used:

$$d_k = (1/14) * \sum_{b=2}^{15} \text{diff}_{\{k,b\}}$$

The table below shows, for each step: the full 16-element diff array (for reference), the extracted 14-element sub-array, its sum, and d_k .

Step	Full diff (b0–b15)	Sub-array b2–b15	Sum	/14	d_k
100	[401, 74, 29, 17, 9, 5, 5, 4, 3, 3, 2, 1, 1, 1, 1, 0]	[29, 17, 9, 5, 5, 4, 3, 3, 2, 1, 1, 1, 1, 0]	81	/14	5.7857
200	[362, 76, 28, 15, 7, 5, 3, 2, 0, -1, -2, -2, -3, -2, -6, -5]	[28, 15, 7, 5, 3, 2, 0, -1, -2, -2, -3, -2, -6, -5]	39	/14	2.7857
300	[377, 77, 26, 13, 5, 1, -2, -6, -8, -10, -13, -16, -18, -19, -18, -23]	[26, 13, 5, 1, -2, -6, -8, -10, -13, -16, -18, -19, -18, -23]	-88	/14	-6.2857
400	[352, 71, 16, 1, -4, -10, -12, -22, -22, -31, -34, -41, -43, -42, -49, -38]	[16, 1, -4, -10, -12, -22, -22, -31, -34, -41, -43, -42, -49, -38]	-331	/14	-23.6429
500	[313, 70, -3, -18, -21, -25, -37, -44, -54, -56, -50, -57, -49, -46, -38, -36]	[-3, -18, -21, -25, -37, -44, -54, -56, -50, -57, -49, -46, -38, -36]	-534	/14	-38.1429
600	[316, 75, 3, -17, -20, -28, -41, -49, -48, -50, -51, -53, -48, -40, -39, -36]	[3, -17, -20, -28, -41, -49, -48, -50, -51, -53, -48, -40, -39, -36]	-517	/14	-36.9286
700	[281, 82, 22, -9, -10, -23, -28, -43, -47, -44, -45, -48, -42, -44, -31, -36]	[22, -9, -10, -23, -28, -43, -47, -44, -45, -48, -42, -44, -31, -36]	-428	/14	-30.5714

Figure 11: t-Test Report (Page 1). The figure Gemma4-E4B Deep Intrinsic Dimensionality Collapse is provided in Figure 7

Step	Full diff (b0-b15)	Sub-array b2-b15	Sum	/14	d_k
800	[251, 67, 21, -18, -11, -22, -30, -46, -40, -43, -48, -53, -45, -40, -31, -38]	[21, -18, -11, -22, -30, -46, -40, -43, -48, -53, -45, -40, -31, -38]	-444	/14	-31.7143
900	[228, 64, 23, -14, -11, -22, -31, -41, -41, -39, -39, -49, -37, -40, -31, -33]	[23, -14, -11, -22, -31, -41, -41, -39, -39, -49, -37, -40, -31, -33]	-405	/14	-28.9286
1000	[210, 56, 35, -15, -7, -19, -29, -38, -40, -38, -37, -46, -42, -41, -34, -34]	[35, -15, -7, -19, -29, -38, -40, -38, -37, -46, -42, -41, -34, -34]	-385	/14	-27.5000
Total/ Mean			-3012		-21.5143

Table 1. Full diff arrays (reference only) and block 2–15 sub-arrays for all 10 checkpoints.

Step 2 — One-Sample t-Test Computation (Left-Tailed, Blocks 2–15)

The 10 values {d_k} extracted from blocks 2–15 are now treated as a random sample. We test whether their population mean mu_d is significantly negative.

Step	Formula	Result
1 Sub-array d_k	$d_k = (1/14) * \text{sum diff}_{[k,b]} \text{ for } b \text{ in } 2..15$	s100: 5.786 s200: 2.786 s300: -6.286 s400: -23.643 s500: -38.143 s600: -36.929 s700: -30.571 s800: -31.714 s900: -28.929 s1000: -27.500
2 Overall mean d_bar	$d_bar = (1/n) * \text{sum } d_k, n = 10$	-21.514286
3 Sample variance s^2	$s^2 = \text{sum}(d_k - d_bar)^2 / (n-1)$	262.578005
4 Sample std dev s_d	$s_d = \text{sqrt}(s^2)$	16.204259
5 Standard error SE	$SE = s_d / \text{sqrt}(n) = s_d / \text{sqrt}(10)$	5.124237
6 t statistic	$t = (d_bar - 0) / SE$	-4.198535
7 Degrees of freedom	$df = n - 1 = 9$	9
8 Critical value (left tail)	$t_{[0.05, df=9]} \text{ (from t-table)}$	-1.8331
9 Decision rule	Reject H0 if $t < t_crit \text{ (left-tail)}$	-4.1985 < -1.8331 => Reject H0
10 p-value (left tail)	$P(T_9 \leq t \mid H0: \mu=0) \text{ via numerical integration}$	p 0.001156

Table 2. Step-by-step t-test computation for blocks 2–15.

Sample mean d_bar	Std dev s_d	Std error SE	t statistic
-21.5143	16.2043	5.1242	-4.1985

CONCLUSION (Blocks 2–15): t = -4.1985 << t_crit = -1.8331 | p-value 0.001156 << 0.05 We strongly reject H0 at alpha = 0.05 (left-tailed, df = 9). d_bar = -21.5143: the output ID in deep layers (blocks 2–15) is systematically and significantly lower than the input ID, revealing a clear dimension-collapse pattern that was masked by blocks 0–1 in the full-range test.

Step 3 — Visualization & Interpretation

3.1 Per-Step d_k Values (Blocks 2–15) with Variance Decomposition

The sign of d_k reverses during training: positive at Steps 100–200 (the residual influence of early, near-block-2 positive values), then strongly negative from Step 300 onward. The table below provides the full variance decomposition.

Exp. k	Step	d_k (b2-15 mean)	$(d_k - \bar{d})$	$(d_k - \bar{d})^2$
1	100	5.785714	+27.300000	745.2900
2	200	2.785714	+24.300000	590.4900
3	300	-6.285714	+15.228571	231.9094
4	400	-23.642857	-2.128571	4.5308
5	500	-38.142857	-16.628571	276.5094
6	600	-36.928571	-15.414286	237.6002
7	700	-30.571429	-9.057143	82.0318
8	800	-31.714286	-10.200000	104.0400
9	900	-28.928571	-7.414286	54.9716
10	1000	-27.500000	-5.985714	35.8288
Sum/Mean		-21.514286	0.000000	2363.2020

Table 3. d_k values (blocks 2–15 only) and deviation terms.

3.2 Temporal Pattern of Deep-Layer Collapse

The block 2–15 data reveal a clear training-phase structure:

- **Early phase (Steps 100–200, $d_k > 0$):** The first few elements of the sub-array (block 2 and 3) are still positive (+29/+17 at Step 100, +28/+15 at Step 200), pulling the 14-element mean slightly positive. Deep layers (blocks 8–15) are near zero or mildly negative. Dimension compression has not yet propagated into the network.
- **Transition (Step 300, $d_k = -6.29$):** The positive entries at blocks 2–4 shrink rapidly while blocks 7–15 turn strongly negative (reaching -23 at block 15). This marks the onset of systematic deep-layer compression.
- **Peak collapse (Steps 400–600, d_k down to -38):** The sub-array is almost entirely negative. Block 2 remains slightly positive or near zero, but blocks 5–15 show values of -10 to -57, indicating aggressive dimensionality reduction throughout the deep network.
- **Stabilisation (Steps 700–1000, d_k around -28 to -31):** The compression magnitude stabilises. Block 2 shows small positive values (+21 to +35) while blocks 4–15 settle into a consistent range of -7 to -46, suggesting an equilibrium between expansion near the output of block 1 and compression in the rest of the network.

3.3 Comparison: Full Range vs. Blocks 2–15

Parameter	Full range (b0-b15)	Deep layers (b2-b15)
Blocks analysed	0 – 15 (16 elements)	2 – 15 (14 elements)
$\bar{d}$	4.9437	-21.5143
s_d	17.4581	16.2043

Parameter	Full range (b0-b15)	Deep layers (b2-b15)
SE	5.5207	5.1242
t statistic	0.8955	-4.1985
t_crit (left, a=0.05)	-1.8331	-1.8331
p-value (left tail)	> 0.05 (not sig.)	p 0.001156
Decision	Cannot reject H0	REJECT H0
Reason	Block 0/1 positive offset masks collapse	Collapse signal fully exposed

Table 4. Side-by-side comparison of the full-range and deep-layer t-tests.

3.4 Full Statistical Summary (Blocks 2–15)

Parameter	Value	Interpretation
Analysis scope	Blocks 2–15 (14 per exp.)	Excludes b0/b1 outliers
H0: $\mu_d = 0$	Tested	No deep-layer change
H1: $\mu_d < 0$	Left-tailed	Output ID < Input ID
n	10	One per checkpoint
d_bar	-21.5143	All neg. from step 300
s_d	16.2043	Large spread: phase transition
SE	5.1242	$s_d / \sqrt{10}$
t statistic	-4.1985	Deep in left tail
df	9	n - 1
t_crit (a=0.05)	-1.8331	Left-tail threshold
p-value	p 0.001156	Highly significant
Decision	Reject H0	Collapse confirmed

Table 5. Full statistical summary of the one-sample left-tailed t-test on blocks 2–15.

FINAL VERDICT (Blocks 2–15): $t = -4.1985$, $p = 0.001156$. The intrinsic dimension collapse in Gemma4-E4B's deep layers (blocks 2–15) is statistically highly significant and cannot be attributed to chance. The block 0/1 expansion effect (a separate phenomenon) had concealed this signal in the full-range test. Isolating the deep layers reveals a collapse pattern qualitatively similar to DeepSeek-V2-Lite, emerging from Step 300 and stabilising from Step 700 onward.

The 10 values {d_k} are treated as a random sample. We test whether their population mean mu_d differs significantly from 0 in the negative direction.

Step	Formula	Result
1 Compute d_k	$d_k = (1/16) * \text{sum diff}_{\{k,b\}}$	s100: -0.125 s200: -23.875 s300: -23.188 s400: -23.875 s500: -22.438 s600: -21.625 s700: -22.500 s800: -23.000 s900: -22.438 s1000: -19.938
2 Overall mean d_bar	$d_{\text{bar}} = (1/n) * \text{sum } d_k, n = 10$	-20.300000
3 Sample variance s_d^2	$s_d^2 = \text{sum}(d_k - d_{\text{bar}})^2 / (n-1)$	51.564931
4 Sample std dev s_d	$s_d = \text{sqrt}(s_d^2)$	7.180873
5 Standard error SE	$SE = s_d / \text{sqrt}(n) = s_d / \text{sqrt}(10)$	2.270791
6 t statistic	$t = (d_{\text{bar}} - 0) / SE$	-8.939615
7 Degrees of freedom	$df = n - 1 = 9$	9
8 Critical value (left tail)	$t_{\{0.05, df=9\}}$ (from t-table)	-1.8331
9 Decision rule	Reject H0 if $t < t_{\text{crit}}$	-8.9396 < -1.8331 => Reject H0
10 p-value (left tail)	$P(T_9 \leq t \mid H_0: \mu=0)$ via numerical integration	p < 0.0001

Table 2. Step-by-step t-test computation.

Sample mean d_bar	Std dev s_d	Std error SE	t statistic
-20.3000	7.1809	2.2708	-8.9396

CONCLUSION: $t = -8.9396 \ll t_{\text{crit}} = -1.8331$ | $p\text{-value} < 0.0001 \ll 0.05$ We strongly reject H0 at the $\alpha = 0.05$ level (left-tailed test, $df = 9$). The mean difference $d_{\text{bar}} = -20.3000$ is significantly negative, providing strong statistical evidence of intrinsic dimension collapse throughout training in Qwen.

Step 3 — Visualization & Interpretation

3.1 Per-Step Mean Difference d_k

Nine out of 10 values of d_k are strongly negative. Step 100 yields a near-zero mean (d_k = -0.125) due to the large positive value at block 0 (+471) offsetting the negative contributions from blocks 1–15. The table below lists all d_k alongside the deviation from the grand mean, providing the inputs for the variance calculation.

Experiment k	Step	d_k = mean diff	(d_k - d_bar)	(d_k - d_bar)^2
1	100	-0.125000	+20.175000	407.030625
2	200	-23.875000	-3.575000	12.780625
3	300	-23.187500	-2.887500	8.337656
4	400	-23.875000	-3.575000	12.780625

Figure 12: t-Test Report (Page 2). The figure Qwen2.5 Intrinsic Dimensionality Collapse is provided in Figure 8

Experiment k	Step	d_k = mean diff	(d_k - d_bar)	(d_k - d_bar)^2
5	500	-22.437500	-2.137500	4.568906
6	600	-21.625000	-1.325000	1.755625
7	700	-22.500000	-2.200000	4.840000
8	800	-23.000000	-2.700000	7.290000
9	900	-22.437500	-2.137500	4.568906
10	1000	-19.937500	+0.362500	0.131406
Sum / Mean		-20.300000	0.000000	464.084375

Table 3. Per-step d_k values and deviation terms for variance computation.

3.2 Pattern of Intrinsic Dimension Collapse

The data reveal a clear two-phase structural pattern across training:

- **Anomalous early phase (Step 100):** Block 0 shows an exceptionally large positive diff (+471), the largest single value in the entire dataset, while blocks 1–15 are uniformly negative (roughly -17 to -58). These opposing forces nearly cancel out, producing a near-zero mean d_k = -0.125. This step is a unique outlier and should not obscure the dominant collapse pattern.
- **Transition (Step 200):** Block 0 reverts to a near-zero positive value (+14), while blocks 1–15 remain strongly negative (-6 to -67). The mean d_k drops sharply to -23.875, marking the onset of systematic collapse.
- **Full collapse with block 0 reversal (Steps 300–1000):** Block 0 itself turns sharply negative (-288 to -338), joining the rest of the network in dimensional compression. Blocks 1–15 converge toward small negative or zero values (-1 to -26). The mean d_k stabilises in the range -20 to -24, indicating a consistent and uniform collapse across all 16 blocks.
- **Gradual recovery at block 0 (Steps 400–1000):** The magnitude of block 0's negative diff decreases from -336 at Step 400 to -302 at Step 1000, suggesting a slow partial stabilisation of the first block's output dimensionality as training progresses.

3.3 Statistical Summary

Parameter	Value	Interpretation
H0: $\mu_{u_d} = 0$	Tested	No systematic collapse
H1: $\mu_{u_d} < 0$	Left-tailed	Output ID < Input ID
n	10	One per checkpoint
d_bar	-20.3000	9/10 steps strongly negative
s_d	7.1809	High spread due to Step 100 outlier
SE	2.2708	$s_d / \sqrt{10}$
t statistic	-8.9396	Far into left tail
df	9	n - 1
t_crit (alpha=0.05)	-1.8331	Left-tail threshold
p-value	p < 0.0001	Highly significant
Decision	Reject H0	Collapse confirmed

t-Test Report of Qwen2.5(Page 3)

Table 4. Full statistical summary of the one-sample t-test.

FINAL VERDICT: With $t = -8.9396$ and $p < 0.0001$, we conclude that the intrinsic dimension collapse observed in Qwen is statistically highly significant. The output ID is systematically and consistently lower than the input ID across all 10 training checkpoints. Unlike DeepSeek-V2-Lite-Chat where the collapse concentrates at block 0 from Step 300, Qwen exhibits a two-phase pattern: an anomalous large positive block-0 value at Step 100 followed by a full-network collapse from Step 200 onward where block 0 itself also reverses to strongly negative. This phenomenon cannot be attributed to random variation.