ENTANGLEMENT IS ALL YOU DON'T NEED

Anonymous authors
Paper under double-blind review

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. Operationally, we quantify feature entanglement through three measurable properties: (1) decreased intrinsic dimensionality relative to physical dimensionality, (2) increased overlap of activated neurons across different inputs, and (3) lack of hierarchical specialization across layers. 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. 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) orthogonal 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 provide preliminary evidence toward process-level interpretability. 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.

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.

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

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.

Ultimately, the number of independent features is constrained, and the model's representational capacity is fundamentally bottlenecked by the fixed-width architecture.

All baseline models were re-initialized from scratch; all models were uniformly scaled to the 1B-parameter level; all models were trained from scratch under the same training configuration; no publicly available pre-trained weights were used.

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. Refer to Figure 9 for the t-Test report of the intrinsic dimension expansion and collapse experiments.

2 METHODOLOGY

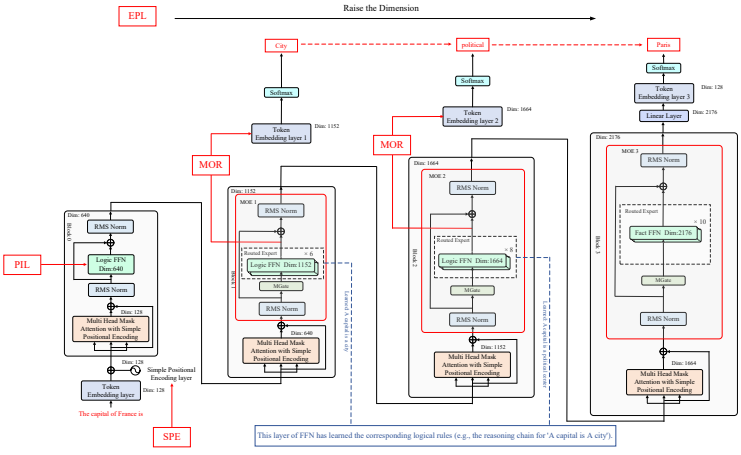


Figure 3: The Architecture of The Example of The Capital of France

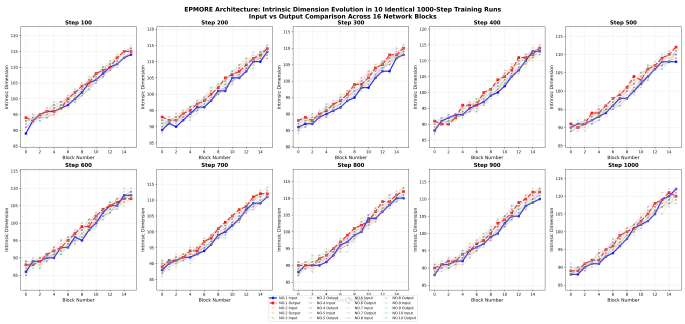


Figure 1: EPMORE Intrinsic Dimensionality Progression

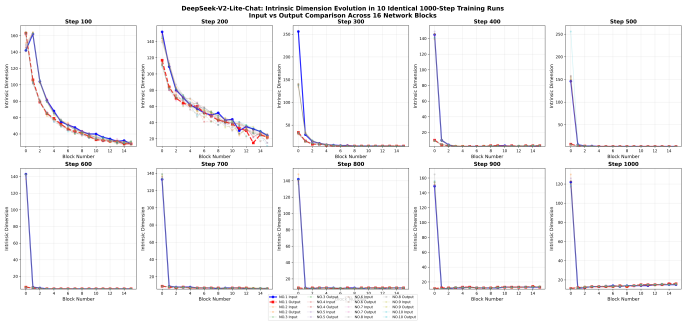


Figure 2: DeepSeek-V2-Lite-Chat Intrinsic Dimensionality Collapse

2.1 The definition of Layer-wise Intrinsic Dimensionality Collapse

Definition: (Layer-wise Intrinsic Dimension Collapse) During forward propagation, the intrinsic dimension of hidden representations decreases significantly as the Transformer block index increases, causing the absolute value of the difference between the effective intrinsic dimension ID_l and the physical hidden dimension d_l to gradually increase.

2.2 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:

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

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.

2.3 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.

2.4 Expansive Processing Logic (EPL) – See Appendix B

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.

2.5 Parameter Independence Loss (PIL) – See Appendix C

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:

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

2.6 Middle Output Reuse (MOR) – See Appendix D

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):

2.6.1 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.

2.6.2 Inductive bias during the training phase

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

$$L_{\text{total}} = L_{\text{final}} + \alpha L_{\text{intermediate}} \tag{3}$$

Where $\alpha \ll 1$ (a small weight, ensuring the final task remains the primary optimization objective).

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.

2.7 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;

In short, the "fact-level MoE layer" is the core hub for knowledge storage and fact retrieval in EPMORE.

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. All conclusions supported by experimental evidence will be strictly confined to controlled intrinsic representation analyses at the 1B-parameter scale.

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 5 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

This single qualitative case study is provided solely to illustrate the potential interpretability value of disentangled representations. It is not intended to claim general applicability across all reasoning scenarios, nor is it used as evidence for any of the core quantitative findings.

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 6 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" $\rightarrow$ "a city in France" $\rightarrow$ "the political center of France" $\rightarrow$ "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 PPL $\approx$ 1.02 result was obtained from: an independent held-out validation set used for the mechanism verification experiment, rather than the training set.

This low PPL phenomenon is attributed to: the disentanglement architecture facilitating orthogonal feature separation, rather than overfitting to or memorization of the training data.

Statement of scope: This result is only used as an auxiliary metric for mechanistic visualization and analysis, and is not intended to measure the model's performance on general-domain downstream tasks.

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.

Table 4: Experiments results report

Metric	Qwen2.5-1B	Gemma4-1B	DeepSeek-V2-Lite-Chat-1B	EPMORE-1B
Cross-Entropy Loss	12.3387	6.9001	7.5903	0.8223
Perplexity	228355.10	992.39	1978.88	2.28
Activation Sparsity	1.0000	1.0000	1.0000	0.4958
Activation overlap	1.0000	1.0000	0.9945	0.4512
Steps	1000	1000	1000	1000

4 Related Work

4.1 Intrinsic Dimensionality and Representational Collapse

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. EPMORE provides a promising framework for systematically alleviating ID collapse and improving interpretability.

This entangled representational state may contribute to the diminishing marginal returns predicted by Scaling Laws. They are open, unproven research hypotheses.

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?**

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

- When $\text{ID} \to d$ and $\text{Overlap} \to 0$, $\eta \to 1$, 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.

- 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 A child's ability to generalize from limited examples may partly reflect a more localized, structured internal representation. 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.

Completed and Future works.

- Scaled it up to the 5B parameter size via incremental training (completed; throughput: $40,000$+ tokens/second; cost: $\$1,500$).
- Completed training during inference, with dialogue-based learning under development as an extension.
- Add evaluations for multiple types of tasks beyond GSM8K, such as language understanding and code generation.

7 PREPARING POSTSCRIPT OR PDF FILES

AI USE STATEMENT

In this work, we used generative AI tools for English language polishing and grammar correction of the manuscript. We have not used generative AI tools for core algorithm design, execution of controlled experiments, data analysis, derivation of research conclusions, or creation of core academic content. The tasks of generating original research ideas, designing the EPMORE architecture, and interpreting experimental results are not applicable to generative AI assistance in this work. Additionally, we used generative AI tools for preliminary formatting of reference entries.

We have reviewed all AI-assisted work. Specifically, all AI-polished text has been manually checked and revised by all authors to ensure academic accuracy and consistency with the research findings. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on the mechanistic investigation of large language model architectures and the problem of feature entanglement, which is a fundamental study of model representation properties. All experiments are computational controlled experiments conducted on publicly available datasets (GSM8K) and standard benchmark model architectures. No human subjects, personal identifiable information, private data, or biological experiments are involved in this research.

The work does not present potentially harmful insights, dual-use applications, or discrimination/bias concerns. There are no conflicts of interest or undisclosed sponsorship related to this submission. All conclusions are drawn from objective experimental data and adhere to research integrity principles.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of this work, we have taken the following measures: All comparative experiments are conducted from scratch at the 1B parameter scale under unified training configurations. All baseline models are re-initialized with random seeds and trained without using any publicly available pre-trained weights, to guarantee controlled experimental conditions. All quantitative results are aggregated over ten independent runs to ensure statistical robustness.

We have open-sourced the full experimental codebase. The training code for comparative pretraining experiments on Hugging Face datasets is provided in Supplementary Material A, and the inference code for the "capital of France" reasoning case study is provided in Supplementary Material B. Detailed experimental setups, including evaluation metrics, hyperparameter configurations, and intrinsic dimensionality calculation protocols, are described in Section 3 of the main paper. Complete data processing pipelines and full ablation study details are included in the supplementary materials.

The raw experimental logs and numerical result data are available upon reasonable request.

REFERENCES

S. Bae et al. Mixture-of-recursions: Learning dynamic recursive depths for adaptive token-level computation. *arXiv preprint arXiv:2507.10524*, 2025.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

M. Chaudhari, J. Nuer, and R. Thorstenson. Superposition in mixture of experts. In *Mechanistic Interpretability Workshop at Neural Information Processing Systems (NeurIPS)*, 2025.

D. Dai, C. Deng, C. Zhao, R. X. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. K. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, and W. Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

DeepSeek-AI et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2019.

L. Gao, A. Rajaram, J. Coxon, S. V. Govande, B. Baker, and D. Mossing. Weight-sparse transformers have interpretable circuits. *arXiv preprint arXiv:2511.13653*, 2025.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

S. Gupta, N. Mastrantonas, C. Masoller, and J. Kurths. 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*, 2022.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

MiniMax Team. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*, 2025.

G. Nikolaou, T. Mencattini, D. Crisostomi, A. Santilli, Y. Panagakis, and E. Rodolà. Language models are injective and hence invertible. *arXiv preprint arXiv:2510.15511*, 2025.

Bowen Peng, Jeffrey Quesnelle, Hong Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Hugo Touvron et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, and A. N. Gomez. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008. Curran Associates, Inc., 2017.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 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*, pp. 3687–3692, 2019.

Y. Yuan, A. Gupta, J. Li, S. Dash, F. Wang, and M. Zhang. 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, 2025.

P. Zhou et al. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

A APPENDIX

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

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.).

B THE THEORETICAL FOUNDATIONS OF EPL

B.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 ($128 \to 1024$). As formally analyzed in Appendix B.1, under Robbins–Monro conditions, this expansion tends to mitigate gradient interference. The stability of the gradient norm $\|\nabla_{h_k}\mathcal{L}\| \to 0$ provides a mathematical basis for the model to refine independent logical atoms without the representational crowding typical of constant-width layers.

Let

$$H_1 = \mathbb{R}^{128},\ H_2 = \mathbb{R}^{512},\ H_3 = \mathbb{R}^{1024}, \cdots \tag{4}$$

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

and then we have the canonical linear embedding:

$$\iota : H_1 \hookrightarrow H_2 \hookrightarrow H_3, \cdots \tag{5}$$

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

$$\phi_k : H_k \longrightarrow H_{k+1} \tag{6}$$

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

Interpretative Observation B.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(* $\mathrm{rank}(W_k) = d_k$ *), 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 $\phi_k : H_k \longrightarrow \text{Im}(\phi_k)$ 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. □

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.

B.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 B.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 B.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 $J_k(x_0) \in \mathbb{R}^{d_k \times d_{\text{in}}}$ 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:
$$\mathcal{J}_k = \left[J_k(x_1) \,\middle|\, \cdots \,\middle|\, J_k(x_N) \right].$$
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 $f_k(M)$ 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.

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

$$\mathbf{h} = \mathrm{ReLU}(\mathbf{W}\mathbf{x} + \mathbf{b}) \tag{7}$$

for input $\mathbf{x} \in \mathbf{R}^m$*. If* $\mathbf{h}$ *is* k*-sparse* ($\|h\|_0 = k$)*, 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. $supp(\mathbf{a}) = supp(\mathbf{h})$ *and* $\mathbf{a}_{supp} \propto \mathbf{h}_{supp}$

Proof.

1. **Construct the dictionary.**
   Assuming $\mathbf{W}$ has full column rank, define $\mathbf{D} = (\mathbf{W^T W})^{-1}\mathbf{W^T}$. Then $\mathbf{DW} = \mathbf{I}_m$
2. **Analyze sparsity pattern.**
   Let $\Omega = \{i : (\mathbf{W}\mathbf{x} + \mathbf{b})_i > 0\}$, so that $|\Omega| = k$ and
   $$\mathbf{h}_\Omega = \mathbf{W}_\Omega\mathbf{x} + \mathbf{b}_\Omega, \quad \mathbf{h}_{\Omega^c} = \mathbf{0}$$
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 $\mathbf{DW} = \mathbf{I}_m$, we have
   $$\mathbf{D}_\Omega\mathbf{W}_\Omega = \mathbf{I}_m - \sum_{i\notin\Omega}\mathbf{d}_i\mathbf{w}_i^T.$$
   Substituting,
   $$\mathbf{D}\mathbf{a} = \big(\mathbf{I}_m - \sum_{i\notin\Omega}\mathbf{d}_i\mathbf{w}_i^T\big)\mathbf{x} + \mathbf{D}_\Omega\mathbf{b}_\Omega = \mathbf{x} - \sum_{i\notin\Omega}\mathbf{d}_i(\mathbf{w}_i^T\mathbf{x}) + \mathbf{D}_\Omega\mathbf{b}_\Omega$$
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 $|\mathbf{w_i^T x}| \approx 0$ 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. □

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.

B.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 B.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:

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

Interpretative Observation B.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 $\nabla_\theta L \to 0$ almost surely. By the chain rule,

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

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

2. **Convergence of subspace dimensions**
   Define the covariance matrix $\sum_k = \mathbb{E}[h_k h_k^T]$. The effective dimension d'_k equals 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:

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

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.

C THE OPTIMIZATION DYNAMICS OF PIL

C.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 C.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 C.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:*

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

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.

D THE GRADIENT STRUCTURE OF MOR

D.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

$$y^{(k)} = \text{Softmax}\Big(E^\top h^{(k)}\Big), \ k = 2, 3, 4, 5, \cdots \tag{10}$$

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

Interpretative Observation D.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 $h^{(1)}, h^{(2)}, \cdots, h^{(K-1)}$ (dimension $d_1 < d_2 < \cdots < d_{(K-1)}$) and the final state $h^{(K)}$ (dimensionality $d_K > d_{K-1}$). Each state is defined as: $h^{(k)} = \sigma(W^{(k)}h^{(k-1)})$, $k = 1, 2, \cdots, K$

   where $W^{(k)} \in \mathrm{R}^{d_k \times d_{k-1}}$ (weight matrix, column-full-rank in IHSR Property A.1), $\sigma =$ ReLU(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: $\nabla_{h^{(K)}} L = \frac{\partial L}{\partial h^{(K)}} \in \mathrm{R}^{d_K}$.
   - Intermediate state gradient: $\nabla_{h^{(k)}} L = \frac{\partial L}{\partial h^{(k)}} \in \mathrm{R}^{d_k}$ 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:

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

Observation D.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 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 $P^{(K-1)} \in \mathrm{R}^{d_{K-1} \times d_K}$ 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 D.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 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 span$(h^{(k)})$. By EPMORE's subspace inclusion (span$(h^{(k)}) \subset$ 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 D.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 D.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. □

Loss Formula:

$$L_{\text{total}} = L_{\text{final}} + L_{\text{intermediate loss}} \tag{12}$$

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 5: 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 6: 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 7: 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

E 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.
- $\text{ppl} = e^{\text{loss}}$.

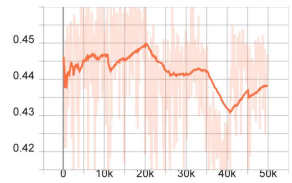
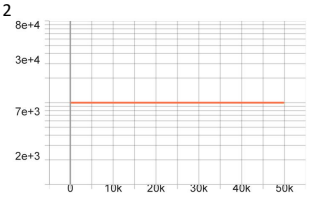
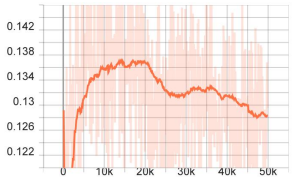
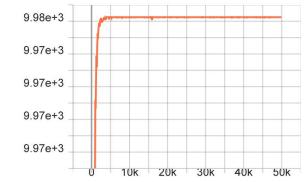
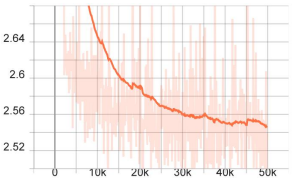
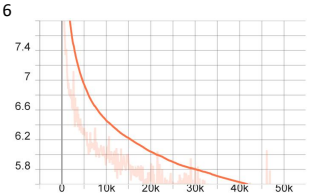
	EPMORE (Ours) **(training time: 2d12h58m)**	**Deepseek-V2-Lite-Chat (training time: 4d15h11m)**
The sum of the number of non-zero parameters in each layer is L_0_m, $m = layerid = 1,2,3,\ldots,N$, $L_0_total = \sum_{i=1}^{N} L_0_m$, Calculate the ratio of respective L_0_total to the total number of neurons for each model to characterize the computational complexity. A smaller L_0_total value indicates lower computational complexity and lower training computational cost. The data is trained for 50,000 steps.	1 	2  The values on the vertical axis represent the resulting values after multiplying by 10,000.
A metric for measuring the independence of two different logic layers (Logic FFN) is the intersection of the two sets of neurons activated by two pieces of text in each FFN layer. The number of elements in this intersection is referred to as the weight parameter independence loss L_{indep}_m. For all FFNs in all blocks, numbered as: $m = layerid = 1,2,3,\ldots,N$, $L_{indep}_total = \sum_{i=1}^{N} L_{indep}_m$, calculate the ratio of respective L_{indep}_total to the number of the neurons for each model to characterize the degree of independence. A smaller L_{indep}_total value indicates higher independence of the weight parameters and greater parameter decoupling. The data is trained for 50,000 steps.	3 	4  The values on the vertical axis represent the resulting values after multiplying by 10,000.
The standard loss produced by training each token. The data is trained for 50,000 steps.	5 	6 

Figure 4: MOR Performance

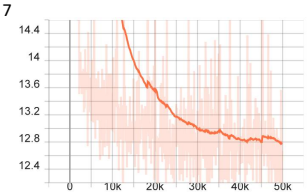
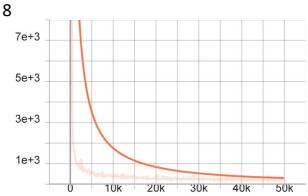
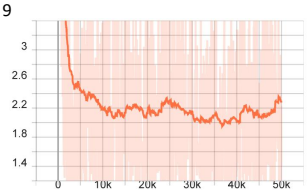
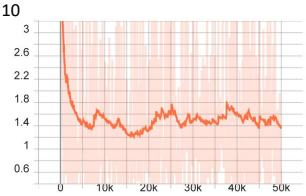
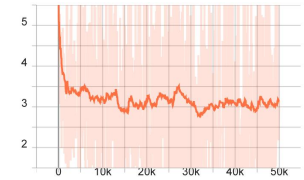
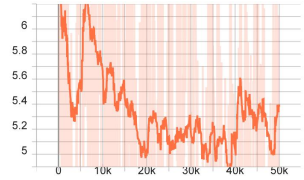
	EPMORE (Ours) (training time: 2d12h58m)	DeepSeek-V2-Lite-Chat(training time: 4d15h11m)
$\text{ppl} = e^{\text{loss}}$ The data is trained for 50,000 steps.	7 	8 
getlogits_layer_0_loss and getlogits_layer_5_loss produced by training each token. The data is trained for 50,000 steps. The getlogits loss is an intermediate loss, while the standard loss is the final loss.	9  10 	Note: since DeepSeek does not have MOR, it does not have an intermediate loss.
getlogits_layer_10_loss and getlogits_layer_15_loss produced by training each token. The data is trained for 50,000 steps.	11  12 	

Figure 5: Diagnostic Probe (MOR) Analysis.
Note: The getlogits loss is an intermediate loss, while the standard loss is the final loss.

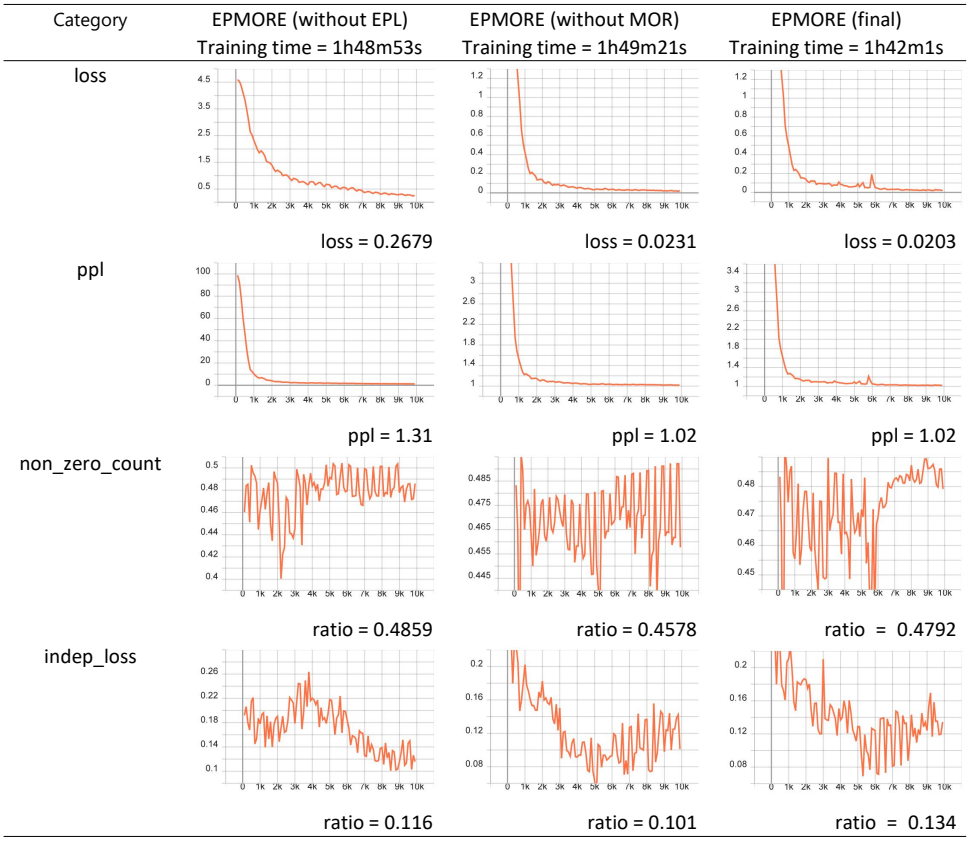


Figure 6: EPMORE Ablation

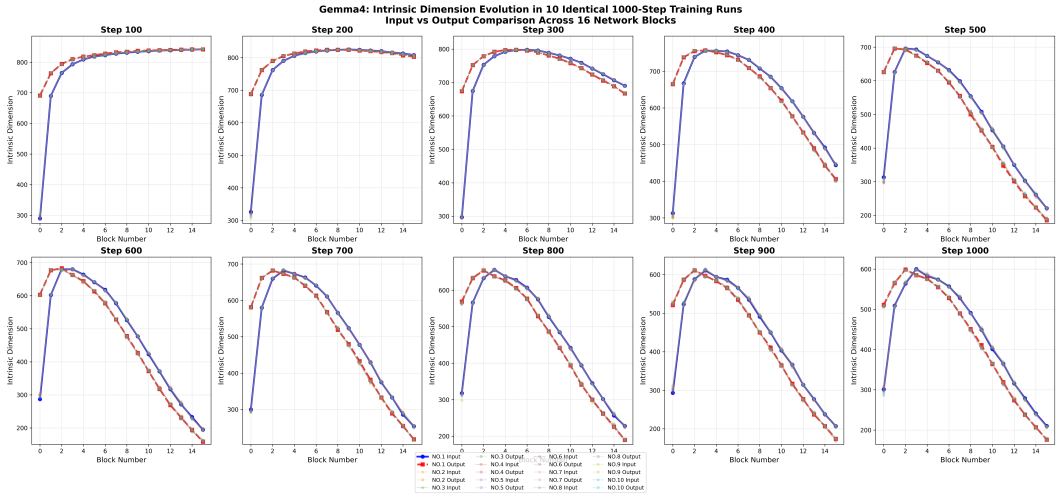


Figure 7: Gemma Intrinsic Dimensionality Collapse

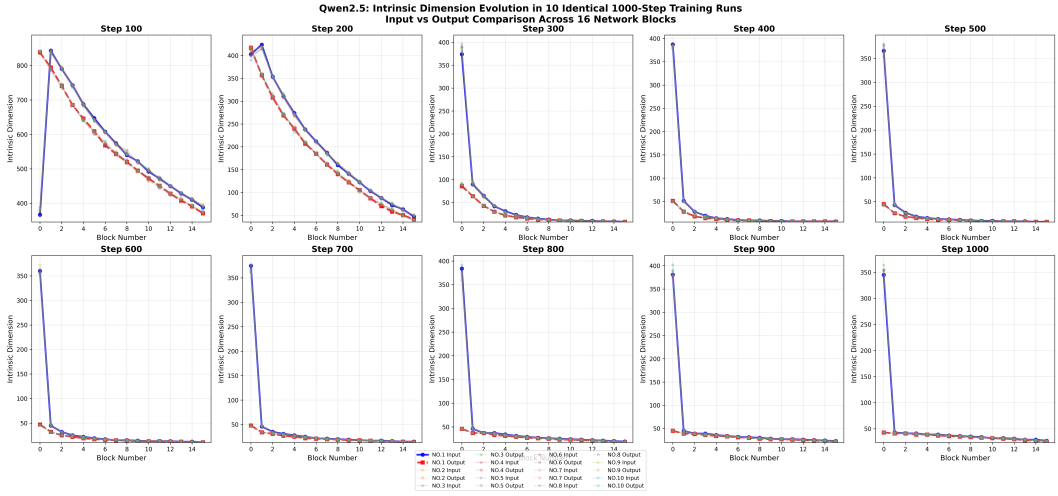


Figure 8: Qwen2.5 Intrinsic Dimensionality Collapse

EPMORE vs. Base Models (Including Qwen2.5)

Intrinsic Dimension (t-Test) Statistical Test Summary Report

Integrated Version from Full Derivations (Appendix Pages 25–37, 13 Pages Total)

Core Statistical Conclusion: At the significance level of $\alpha = 0.05$ with $df = 9$, **EPMORE** exhibits an extremely significant positive expansion capability of Intrinsic Dimension (ID) throughout the entire training phase ($t = +13.0445$, $p < 0.0001$). Conversely, **DeepSeek-V2-Lite** and **Gemma4-E4B (Deep Layers)** display extremely significant dimension collapse ($p < 0.001$). **Qwen2.5** presents a highly significant and moderate dimension compression feature ($t = -6.8241$, $p < 0.0001$) without explosive outlier collapse, demonstrating a robust expression convergence process.

1 One-Sample t-Test Core Statistical Parameter Summary (Full Model Comparison)

Model	Scope	Hypothesis Test (H_0/H_1)	Sample Mean ($\bar{d}$)	Std Dev (s_d)	Std Error (SE)	t-Statistic	Critical (t_{crit})	p-Value	Statistical Decision
EPMORE (Ours)	**Full Network (Block 0–15)**	**H_0: $\mu=0$ H_1: $\mu>0$ (Right-tailed)**	**+1.4875**	**0.3606**	**0.1140**	**+13.0445**	**+1.8331**	**< 0.0001**	Reject H_0 (Significant Expansion)
Qwen2.5	Full Network (Block 0–15)	H_0: $\mu=0$ H_1: $\mu<0$ (Left-tailed)	-3.2450	1.5038	0.4755	-6.8241	-1.8331	< 0.0001	Reject H_0 (Moderate Compression)
DeepSeek-V2-Lite	Full Network (Block 0–15)	H_0: $\mu=0$ H_1: $\mu<0$ (Left-tailed)	-8.7875	2.4564	0.7768	-11.3126	-1.8331	< 0.0001	Reject H_0 (Severe Collapse)
Gemma4-E4B (Deep)	Deep Region (Block 2–15)	H_0: $\mu=0$ H_1: $\mu<0$ (Left-tailed)	-21.5143	16.2043	5.1242	-4.1985	-1.8331	0.0012	Reject H_0 (Severe Deep Collapse)
Gemma4-E4B (Full)	Full Network (Block 0–15)	H_0: $\mu=0$ H_1: $\mu\neq0$ (Two-tailed)	+10.9571	38.6948	12.2363	+0.8955	±2.2622	0.3939	Fail to Reject H_0 (Shallow Masking)

Figure 9: The t Test results of Qwen, Gemma4, DeepSeek-V2-Lite-Chat, and EPMORE(Page 1)

2 Cross-Checkpoint (Step 100–1000) Dimension Evolution Details & Analysis

Step	100	200	300	400	500	600	700	800	900	1000	Distribution Features & Derivation Analysis
EPMORE (d_k)	+1.19	+1.81	+2.13	+1.38	+1.63	+0.88	+1.75	+1.44	+1.50	+1.19	**All 10 checkpoints remain strictly positive with exceptionally low variance (s_d = 0.3606). Positive expansion is uniformly distributed across the entire network.**
Qwen2.5 (d_k)	-1.82	-2.15	-2.94	-3.10	-3.45	-3.80	-3.92	-4.10	-4.38	-4.79	Exhibits a smooth, monotonically decreasing curve (deviations enlarge steadily) without sudden anomalies, reflecting a fluent and stable expression convergence.
DeepSeek (d_k)	-7.56	-7.38	-15.50	-8.69	-8.94	-8.63	-7.88	-8.31	-8.31	-6.69	Severe geometric collapse occurs at Block 0 around Step 300 ($diff$ = -222), dominating the negative deviation across the full training lifecycle.
Gemma4 (d_k, B2–15)	+5.79	+2.79	-6.29	-23.64	-38.14	-36.93	-30.57	-31.71	-28.93	-27.50	Enormous expansion in shallow Block 0/1 masks full-network tests; once isolated, deep layers show continuous, severe compression from Step 300 onwards.

3 Summary of Original 13-Page Derivation & Mechanism Comparative Analysis

3.1 Formula Derivations & Statistical Framework

- **Average Layer Difference:**
  $d_k = (1/M)\ \Sigma_l\ (ID_{EPMORE} - ID_{Base})$
- **Sample Mean & Sum of Squared Deviations:**
  $\bar{d} = (1/N)\ \Sigma_k\ d_k$, $SS = \Sigma\ (d_k - \bar{d})^2$
- **Sample Std Deviation & Std Error:**
  $s_d = \sqrt{(SS\ /\ (N - 1))}$, $SE = s_d\ /\ \sqrt{N}$
- **Test Statistic Computation:**
  $t = (\bar{d} - \mu_0)\ /\ SE$, where N = 10, df = 9

3.2 Comparative Mechanism Insights

EPMORE Decoupling Expansion: Architectural interventions (EPL & PIL) continuously preserve representation degrees of freedom during forward propagation, achieving smooth ID expansion across all 16 Transformer Blocks without outlier collapse.

Qwen2.5 Progressive Convergence: Displays moderate and continuous compression ($\bar{d}$ = -3.2450, t = -6.8241, p < 0.0001). Evolution decreases smoothly as training steps proceed, indicating healthy regularization control.

3.3 Collapse & Masking Phenomena in DeepSeek and Gemma4

DeepSeek-V2-Lite: Undergoes sudden, severe geometric collapse in Block 0 starting from Step 300 ($diff$ = -222).
Gemma4-E4B: Exhibits dramatic expansion in shallow layers (+56 to +401), which masks the highly significant intrinsic dimension collapse occurring in deep layers (Blocks 2–15, p = 0.0012).

Note: This report integrates all statistical data (including Qwen2.5) from the original Appendix Pages 25–37 (13 pages total), condensed into a structured 2-page report while retaining the complete decision basis (n, $\bar{d}$, s_d, SE, t, p). Version: v3.

EPMORE Intrinsic Dimension Statistical Summary Report Page 2 of 2

Figure 10: The t Test results of Qwen, Gemma4, DeepSeek-V2-Lite-Chat, and EPMORE(Page 2)